

EUROPEAN COMMISSION  
DIRECTORATE GENERAL  
JOINT RESEARCH CENTRE

Institute for Health and Consumer Protection  
Toxicology and Chemical Substances Unit  
European Chemicals Bureau  
I-21020 Ispra (VA) Italy

# **THE CHARACTERISATION OF (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIPS: PRELIMINARY GUIDANCE**

*Worth AP, Bassan A, Gallegos A, Netzeva TI, Patlewicz G,  
Pavan M, Tsakovska I & Vracko M*

**2005**

**EUR 21866 EN**

## **LEGAL NOTICE**

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of the following information.

A great deal of additional information on the European Union is available on the Internet. It can be accessed through the Europa server (<http://europa.eu.int>)

EUR 21866 EN  
© European Communities, 2005  
Reproduction is authorised provided the source is acknowledged  
*Printed in Italy*

## Abstract

In November 2004, the OECD Member Countries and the European Commission adopted five principles for the validation of (quantitative) structure-activity relationships ([Q]SARs) intended for use in the regulatory assessment of chemicals. International agreement on a set of validation principles was important, not only to provide regulatory bodies with a scientific basis for making decisions on the acceptability of data generated by (Q)SARs, but also to promote the mutual acceptance of (Q)SAR models by improving the transparency and consistency of (Q)SAR reporting.

According to the OECD Principles for (Q)SAR validation, a (Q)SAR model that is proposed for regulatory use should be associated with five types of information: 1) a defined endpoint; 2) an unambiguous algorithm; 3) a defined domain of applicability; 4) appropriate measures of goodness-of-fit, robustness and predictivity; and 5) a mechanistic interpretation, if possible. Taken together, these five principles form the basis of a conceptual framework for characterising (Q)SAR models, and of reporting formats for describing the model characteristics in a transparent manner.

Under the proposed REACH legislation in the EU, there are provisions for the use of estimated data generated by (Q)SARs, both as a substitute for experimental data, and as a supplement to experimental data in weight-of-evidence approaches. It is foreseen that (Q)SARs will be used for the three main regulatory goals of hazard assessment, risk assessment and PBT/vPvB assessment. In the Registration process under REACH, the registrant will be able to use (Q)SAR data in the registration dossier provided that adequate documentation is provided to argue for the validity of the model(s) used.

This report provides preliminary guidance on how to characterise (Q)SARs according to the OECD validation principles. It is emphasised that the understanding of how to characterise (Q)SAR models is evolving, and that the content of the current report reflects the understanding and perspectives of the authors at the time of writing (November 2005). It is therefore likely that an update will be produced in the future for the benefit of those who need to submit (Industry) or evaluate (Authorities) chemical information based (partly) on (Q)SARs. It is also noted that this document does not provide guidance on the use of (Q)SAR reporting formats, or on criteria for the acceptance of (Q)SAR estimates, since EU guidance on these topics stills need to be developed.

## Contents

	<b>Page</b>
Abbreviations	1 - 2
Chapter 1. Introduction to document	3 - 11
Chapter 2. Defined endpoint and algorithm	12 - 16
Chapter 3. (Q)SAR applicability domain	17 - 28
Chapter 4. Statistical validation	29 - 53
Chapter 5. Mechanistic relevance	54 - 67
References	68 - 80
Appendix 1. Check list for the application of the OECD validation principles	81 - 85

## Abbreviations

AD	Applicability Domain
AI	Artificial Intelligence
ANN	Artificial Neural Network
CEFIC	European Chemical Industry Council
CI	Confidence Intervals
CM	Classification Model
CT	Classification Tree analysis
DEREK	Deductive Estimation of Risk from Existing Knowledge
DModX	Distance to model in X space
DModY	Distance to model in Y space
ECETOC	European Centre for Ecotoxicology and Toxicology
ECM	Embedded Cluster Modelling
EMS	Explained Mean Square
FFD	Fractional Factorial Design
GA	Genetic Algorithm
GHS	Globally Harmonised System (for the classification of chemicals)
LUMO	Lowest unoccupied molecular orbital
HOMO	Highest occupied molecular orbital
ICCA	International Council of Chemical Associations
JRC	Joint Research Centre
KNN	K-Nearest Neighbour
LC <sub>50</sub>	Test concentration causing 50% lethality
Log K <sub>ow</sub>	Logarithm of the Octanol-Water Partition Coefficient
Log P	Logarithm of a partition coefficient, e.g. Log (octanol/water)
LOO	Leave-One-Out cross validation technique
LMO	Leave-Many-Out cross validation technique
LR	Logistic Regression
LUMO	Lowest Unoccupied Molecular Orbital
MDA	Multivariate Discriminant Analysis
MLR	Multiple Linear Regression
MSE	Mean Squared Error
MOA	Mechanism (Mode) of (Toxic) Action
MR	Molar Refractivity
MW	Molecular Weight
NN	Neural Networks
OECD	Organisation for Economic Cooperation and Development
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
Pi (p)	Hydrophobicity substituent constant
PLS	Partial Least Squares Projections to Latent Structures
PRESS	PRedictive Error Sum of Squares
(Q)SAR	(Quantitative) Structure-Activity Relationship
(Q)SBR	(Quantitative) Structure-Biodegradability Relationship
(Q)SPR	(Quantitative) Structure Property Relationship
RAI	Relative Alkylation Index
R <sup>2</sup>	Multiple correlation coefficient
REACH	Registration, Evaluation, and Authorisation of Chemicals
RMS	Residual Mean Square

ROC	Receiver Operating Characteristic
<i>s</i>	Standard error of the estimate
sigma ( <i>s</i> )	Electronic substituent constant
SDEP	Standard Deviation Error of Prediction
SIDS	Screening Information Data Set
SMILES	Simplified Molecular Input Line entry system
SN2	(Bimolecular) Nucleophilic Substitution
ULR	Univariate Linear Regression
V <sub>m</sub>	Molar Volume

# **CHAPTER 1**

## **INTRODUCTION**

## CHAPTER 1: INTRODUCTION

### Summary of chapter 1

This chapter provides an introduction to the OECD Guidance Document on (Q)SAR Validation, and makes reference to related OECD documents, including the Guidance Document 34 on the Validation and Acceptance of Test Methods (para 11), and the Guidance Document on the Regulatory Application of (Q)SARs (para 39). The chapter starts with the historical background to the establishment of the OECD Principles for (Q)SAR Validation and to the development of this document (paras 1-10). The concept of validation, as it applies to (Q)SARs, is then defined (paras 11-16), and the intended outcome of a (Q)SAR validation exercise is described (paras 17-19). The five validation principles are then presented (para 20), along with an explanation of the intent of these principles (paras 21-26). The intended coverage of the principles is explained (paras 27-28), including the restriction of the principles, and this guidance document, to the validation of models rather than software packages (para 29-30). The purpose of this guidance document is clarified in (paras 31-33), along with the target audience (paras 34-36). Since the validation of (Q)SARs provides a basis for their regulatory application, some comments are provided on the boundary between validation and regulatory acceptance, which has implications for the limits of this guidance document (paras 37-40). Finally, this chapter concludes with a brief overview of the rest of this document (paras 41-46).

### Historical background

1. A set of six principles for assessing the validity of (Q)SARs were proposed at an international workshop on the “Regulatory Acceptance of QSARs for Human Health and Environment Endpoints”, organised by the International Council of Chemical Associations (ICCA) and the European Chemical Industry Council (CEFIC), and held in Setubal, Portugal, on 4-6 March, 2002 (1,2,3,4).
2. The regulatory use of structure-activity relationships (SARs) and quantitative structure-activity relationships (QSARs), collectively referred to as (Q)SARs, varies considerably among OECD Member Countries, and even between different agencies within the same Member Country. This is partly due to different regulatory frameworks, which impose different requirements and work under different constraints, but also because an internationally harmonised conceptual framework for assessing (Q)SARs has been lacking. The lack of such a framework led to the widespread recognition of the need for an internationally-agreed set of principles for (Q)SAR validation. The development of a set of agreed principles was considered important, not only to provide regulatory bodies with a scientific basis for making decisions on the acceptability (or otherwise) of data generated by (Q)SARs, but also to promote the mutual acceptance of (Q)SAR models by improving the transparency and consistency of (Q)SAR reporting .
3. In November 2002, the 34th Joint Meeting (JM) of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology agreed to start a new OECD activity aimed at increasing the regulatory acceptance of (Q)SARs, and to establish an Expert Group for this work.



4. The 1<sup>st</sup> Meeting of the Expert Group was hosted by the European Commission's Joint Research Centre (JRC), in Ispra, Italy, on 31 March – 2 April, 2003. Following the request of the 34th JM, the participants of the 1<sup>st</sup> Expert Group Meeting proposed a (two-year) work plan for the OECD work on (Q)SARs. The work plan included three Work Items. The aim of Work Item 1, completed in 2004, was to apply the Setubal principles to selected (Q)SARs, in order to evaluate the principles, and to refine them wherever necessary. The aim of Work Item 2 was to develop guidance documents for the validation of (Q)SARs to assist (Q)SAR practitioners and (Q)SAR end-users in developing and evaluating (Q)SARs with respect to the validation principles. The aim of Work Item 3 was to identify practical approaches to make (Q)SARs readily available and accessible to scientists in regulatory bodies, industry and universities.

5. To manage the OECD work plan on QSARs, the 1<sup>st</sup> Expert Group Meeting proposed a subgroup, called the Coordinating Group of the Expert Group on (Q)SARs. In June 2003, the proposed work plan was endorsed by the 35th JM. At the same meeting, the European Commission (JRC) offered to take the lead in coordinating Work Item 1 on the evaluation of the Setubal Principles, with the support of the Coordinating Group. The offer was welcomed and accepted by the 35th JM.

6. To carry out Work Item 1, a team of experts (the Work Item 1 Team) produced a total of eleven case studies, by applying the Setubal principles to specific (Q)SARs or software models. The models chosen included literature-based models for acute fish toxicity, atmospheric degradation, mutagenicity and carcinogenicity, and the following software models: the Multi-CASE model for *in vitro* chromosomal aberrations; Multi-CASE and MDL models for human NOEL; ECOSAR; BIOWIN; DEREK; the DEREK skin sensitisation rulebase; the Japanese METI biodegradation model; and the rat oral chronic toxicity models in TOPKAT. These models were considered to collectively provide a representative range of (Q)SAR approaches, covering a variety of physicochemical, environmental, ecological and human health endpoints.

7. To provide guidance on the application of the proposed principles, a check list of considerations (questions) was developed by the Coordinating Group, and this was refined on the basis of experience obtained by carrying out Work Item 1). The refined check list was presented to the 16th Meeting of the Working Group of National Coordinators of the Test Guidelines Programme, held on 26-28 May 2004.

8. The report on the outcome of Work Item 1 was discussed by the 2<sup>nd</sup> Expert Group Meeting, hosted by the OECD, in Paris, on 20-21 September 2004. The report consisted of a consolidated report by the Coordinating Group, including a proposal for revision of the Setubal principles, followed by a set of annexes containing the 11 case studies. The Expert Group refined the wording of the consolidated report, which included combining the internal and external validation principles into a single principle (5), which then represented the consensus view of the Expert Group. It was also agreed that the views expressed in the annexes (6) should be regarded as views of the identified authors, and not necessarily the views of the Expert Group.

9. The final report on the outcome of Work Item 1 (5,6), and in particular the proposed OECD Principles for (Q)SAR Validation, were adopted by the 37<sup>th</sup> JM on 17-19 November 2004. The JM supported the Expert Group's proposal that Work Item 1 should be followed up with Work Item 2 in the development of this Guidance Document on (Q)SAR Validation,

which should provide detailed and non-prescriptive guidance to explain and illustrate the application of the OECD Principles for (Q)SAR Validation to different types of models.

10. The 37<sup>th</sup> JM also agreed on some changes in the coordination of the OECD QSAR work programme. In particular, oversight of the (Q)SAR project was assigned to the Task Force on Existing Chemicals, and the name of the (Q)SAR Group, often referred to as the “(Q)SAR Expert Group” was changed to “Ad hoc Group on (Q)SARs”. At the same time, the membership of the Ad hoc Group was re-established. Following receipt of the nominations from the Member Countries, the 38<sup>th</sup> JM on 8-10 June 2005 agreed to replace the Coordinating Group with a smaller Steering Group, consisting of those members of the Ad Hoc Group who are most closely involved in the planning and routine management of the (Q)SAR project.

### **Definition of “validation” and the “validation process” in the context of (Q)SARs**

#### ***Definition of (Q)SAR validation***

11. The guidance for (Q)SARs in the present document is consistent with the general guidance given in OECD Guidance Document 34 on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment (7).

12. According to OECD Guidance Document 34, the term “validation” is defined as follows:

“Validation is the process by which the relevance and reliability of a method are assessed for a particular purpose”.

In this definition, relevance refers to the scientific (mechanistic) basis of the experimental method and to the predictive ability of an associated prediction model (in the case of methods where an extrapolation is made across an endpoint and/or species), whereas reliability refers to the reproducibility of the method, both within and between laboratories, and over time.

13. An adaptation of this definition is needed for (Q)SARs to account for the fact that a (Q)SAR is a derivative of experimental data and not the experimental method itself. The term “relevance” is applicable to (Q)SARs, if this is interpreted as referring to the predictive ability of the model and to the possibility to interpret the model in mechanistic terms. Since a (Q)SAR is a derivative of the experimental data, the mechanistic relevance of the model is logically tied to the corresponding endpoint of the experimental method, irrespective of whether this test method is itself considered to be relevant for any particular purpose. The assessment of (Q)SAR reliability, however, places greater emphasis on the accuracy of the (Q)SAR with many different chemicals than on the reproducibility of the (Q)SAR within and between laboratories. In other words, the assessment of reliability within a specific chemical universe (applicability domain) is emphasised more than the reproducibility of individual endpoint estimations. In contrast, the validation of experimental methods has traditionally placed less emphasis on the importance of the chemical domain when assessing the validity of the results.

14. An adaptation of this definition is needed for (Q)SARs, for the following reasons: a) (Q)SAR models are not test methods, which implies some differences between the validation approaches adopted for (Q)SARs and those adopted for test methods; b) in the (Q)SAR field, the term “reliability” is generally used to reflect “accuracy” or “validity”, rather than

reproducibility; and c) the reproducibility of a (Q)SAR does not normally need to be assessed during validation, because the predictions generated by such a model are not expected to exhibit significant intra-laboratory and inter-laboratory variability, such as that associated with experimental methods.

15. In relation to (Q)SARs, the following adaptation of the traditional definition of validation is proposed:

“The validation of a (Q)SAR is the process by which the performance and mechanistic interpretation of the model are assessed for a particular purpose.”

In this definition, the “performance” of a model refers to its goodness-of-fit, robustness and predictive ability, whereas “purpose” refers to the scientific purpose of the (Q)SAR, as expressed by the defined endpoint and applicability domain. The first part of the definition (performance) refers to “statistical validation”, whereas the second part (mechanistic interpretation) refers to the assignment of physicochemical meaning to the descriptors (where possible) and to the establishment of a hypothesis linking the descriptors with the endpoint.

16. The scientific purpose of a (Q)SAR may or may not have an association with possible regulatory applications. Thus, the purpose of a (Q)SAR could be for predicting a particular endpoint (along a continuous or categorical scale) for a particular class of chemicals, irrespective of whether the endpoint is required by any particular legislation or whether the class of chemicals is contained within a given regulatory inventory.

### ***The (Q)SAR validation process***

17. For the purposes of this guidance document, a “validation process” refers to any exercise in which the OECD principles for (Q)SAR validation are applied to a given model or set of models. It is not implied that the validation process should be carried out by any particular organisation, committee or formal validation body.

18. The outcome of a (Q)SAR validation process should be a dossier providing information on the validity of a (Q)SAR. The information should be obtained by applying the (Q)SAR validation principles, and the dossier should be structured accordingly. For scientific and/or practical reasons, it will not be possible to fulfill all principles for all models of regulatory interest. Therefore, the output of a successful validation exercise is the provision of a dossier that is as complete as possible, given the scientific and practical constraints. The output of a successful validation process does not need to include any opinion on the validity of a model.

19. It follows that each regulatory authority will need to apply flexibility when considering the acceptability of a given (Q)SAR, taking into account the information provided in the (Q)SAR validation dossier, and the needs and constraints of its particular regulatory programme.

### **The OECD Principles for (Q)SAR Validation**

20. On the basis of the Work Item 1 report (5,6), the 37<sup>th</sup> Joint Meeting agreed on the following wording of the OECD Principles for (Q)SAR validation:

“To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:

- 1) a defined endpoint;
- 2) an unambiguous algorithm;
- 3) a defined domain of applicability;
- 4) appropriate measures of goodness-of-fit, robustness and predictivity;
- 5) a mechanistic interpretation, if possible.”

These principles should be read in conjunction with the explanatory notes presented in the following section.

### ***Intent of the (Q)SAR validation principles***

21. The principles for (Q)SAR validation and the associated check list are intended to identify the types of information that are considered useful for the regulatory review of (Q)SARs. Taken together, the principles and check list constitute a conceptual framework to guide the validation of (Q)SARs, but they are not intended to provide criteria for the regulatory acceptance of (Q)SARs. The definition of acceptance criteria, where considered necessary, are the responsibility of individual authorities within the Member Countries.

22. According to Principle 1, a (Q)SAR should be associated with a “defined endpoint”, where endpoint refers to any physicochemical, biological or environmental effect that can be measured and therefore modelled. The intent of this principle is to ensure transparency in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions. Ideally, (Q)SARs should be developed from homogeneous datasets in which the experimental data have been generated by a single protocol. However, this is rarely feasible in practice, and data produced by different protocols are often combined.

23. According to Principle 2, a (Q)SAR should be expressed in the form of an unambiguous algorithm. The intent of this principle is to ensure transparency in the description of the model algorithm. In the case of commercially developed models, this information is not always made publicly available.

24. According to Principle 3, a (Q)SAR should be associated with a “defined domain of applicability”. The need to define an applicability domain expresses the fact that (Q)SARs are reductionist models which are inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions. This principle does not imply that a given model should only be associated with a single applicability domain. As discussed in Chapter 3, the boundaries of the domain can vary according to the method used to define it and the desired trade-off between the breadth of model applicability and the overall reliability of predictions.

25. According to Principle 4, a (Q)SAR should be associated with “appropriate measures of goodness-of-fit, robustness and predictivity.” This principle expresses the need to provide two types of information: a) the internal performance of a model (as represented by goodness-of-fit and robustness), determined by using a training set; and b) the predictivity of a model, determined by using an appropriate test set. As discussed in Chapter 4, there is no

absolute measure of predictivity that is suitable for all purposes, since predictivity can vary according to the statistical methods and parameters used in the assessment.

26. According to Principle 5, a (Q)SAR should be associated with a “mechanistic interpretation”, wherever such an interpretation can be made. Clearly, it is not always possible to provide a mechanistic interpretation of a given (Q)SAR. The intent of this principle is therefore to ensure that there is an assessment of the mechanistic associations between the descriptors used in a model and the endpoint being predicted, and that any association is documented. Where a mechanistic interpretation is possible, it can also form part of the defined applicability domain (Principle 3).

### ***Coverage of the (Q)SAR validation principles***

27. The (Q)SAR validation principles are intended to be applicable to a diverse range of models types including SARs, QSARs, decision trees, neural network models, and “complex models”, such as expert systems, which may be based on the use of multiple models. The guidance provided in this document is intended to reflect this diversity of model types.

28. In the case of “complex models” that are actually based on the use of multiple models, it is important to identify the smallest component that functions independently, and to apply the principles to the individual component. Examples of such models include ECOSAR and DEREK for Windows.

### ***Model validation vs software verification***

29. This guidance document covers the validation of models, but not the verification of computer programmes. It is important to distinguish between models, which are generalisations (based on theory or observation) used to make predictions, and computer programmes, which may be developed to implement models on particular computer platforms. Furthermore, it is important to distinguish between the *validation* of a model, and the *verification* of the software programme that applies the model. A highly predictive model could be regarded as valid, without considering whether the model has been coded correctly in the form of a computer programme. Conversely, a poorly predictive model, which might not be regarded as valid, could be accurately translated into a specific programming language for implementation in a specific software package.

30. In principle, any model could be implemented in a variety of computer platforms, However, in practice, for certain types of models, it may be difficult to separate the model from the platform. This is particularly true of commercially available models, where certain components of the model (e.g. training sets, algorithms) are hidden for proprietary purposes.

### ***Purpose of this guidance document***

31. The purpose of this document is to provide detailed and non-prescriptive guidance that explains and illustrates the application of the validation principles to different types of (Q)SAR models.

32. This document is needed to: a) present a harmonised conceptual framework to guide the conduct of (Q)SAR validation studies and to help regulators who will need to consider the

outcome of such studies; and b) explain and illustrate with examples how the validation principles can be interpreted for different types of models.

33. The establishment of criteria for determining the scientific validity and regulatory acceptability of (Q)SARs is not covered by this document. Where considered necessary, it is expected that such criteria will be established by the appropriate regulatory authorities.

### ***Target audience***

34. This document is aimed primarily at (Q)SAR specialists who need to carry out (Q)SAR validation exercises. Therefore, the guidance will be based on the assumption that the reader is familiar with the basics of (Q)SAR.

35. At a secondary level, the document will be aimed at non-specialist stakeholders of the validation process who will not need to perform validation exercises themselves, but who will need a sufficient conceptual basis for understanding their outcome. For example, regulators may need to assess the outcome of a (Q)SAR validation exercise, without necessarily conducting the exercise themselves.

36. To accommodate both types of readership, each chapter contains a summary, written at a more general level for the non-specialist. Each chapter summary makes reference to specific sections of the chapter where the specialist reader can find more detailed information.

### **Regulatory application of (Q)SARs**

37. As mentioned above, the aim of this document is to provide guidance on how the (Q)SAR validation principles can be applied to different types of models, not to define criteria for the validity or acceptability of (Q)SARs.

38. In cases where the application of the principle is inherently subjective, guidance is provided through the use of examples taken from the (Q)SAR literature. In cases where statistical methods or other approaches are available for applying a principle, the current state-of-the-art is described, and advice is given on the strengths and limitations of different approaches.

39. Flexibility will be needed in the interpretation and application of each principle because ultimately, the proper integration of (Q)SARs into any type of regulatory/decision-making framework depends upon the needs and constraints of the specific regulatory authority. The need for such flexibility is given in a case study by the US EPA (8).

40. A separate document on the regulatory application of (Q)SARs is being developed by the OECD Ad hoc Group on (Q)SARs (9), under the coordination of the US EPA.

## **Overview of this document**

41. Each chapter of this document addresses the application of one or more principles.
42. Chapter 2 provides examples to illustrate how the concepts of “unambiguous algorithm” and “defined endpoint” can be interpreted in relation to different types of (Q)SARs.
43. Chapter 3 describes the current state-of-the-art of statistical methods and other approaches for the assessment of the applicability domains of (Q)SARs.
44. Chapter 4 describes the current state-of-the-art of statistical methods and other approaches for assessing the goodness-of-fit, robustness and predictivity of (Q)SARs, and explains the concepts of internal and external validation. This chapter also illustrates, by means of flow charts, logical sequences of steps that could be taken during the validation of (Q)SARs.
45. Chapter 5 provides examples to illustrate how the concept of “mechanistic interpretation” can be applied to (Q)SARs, where feasible. The mechanistic interpretation of a (Q)SAR includes two considerations: a) the interpretation of the (Q)SAR descriptors and consequently their relevance for the prediction of the endpoint; b) the relevance of the mathematical form of the relationship between the descriptors and the endpoint being modelled.
46. Appendix 1 provides a check list of questions to help in the application of the OECD principles for QSAR validation.





## **CHAPTER 2**

### **DEFINED ENDPOINT AND ALGORITHM**

## CHAPTER 2: DEFINED ENDPOINT AND ALGORITHM

### Summary of chapter 2

This chapter introduces the rationale behind the first two OECD validation principles, according to which a (Q)SAR should be associated with a “defined endpoint” (Principle 1) and with a “unambiguous algorithm” (Principle 2). Guidance is provided on the interpretation of these principles, by describing what constitutes a defined endpoint and an unambiguous algorithm. Following an introduction to the establishment of the principles (paras 1-2), the concept of the defined endpoint is discussed (paras 3-7), followed by the concept of the defined algorithm (paras 8-12) is emphasised that a defined endpoint in the context of test guidelines does not necessarily correspond with a defined endpoint in the context of (Q)SAR development. The need for a defined algorithm is discussed in terms of the elements that are needed for an algorithm to be fully transparent.

### Introduction

1. According to Principle 1, a (Q)SAR should be associated with a “defined endpoint”, where endpoint refers to any physicochemical, biological or environmental effect that can be measured and therefore modelled. The intent of this principle is to ensure transparency in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions. When making a comparison between (Q)SAR predictions with experimental data, it is important to know whether the model was intended to generate the same type of information. It is also important to know whether the experimental data used to develop the model were generated according to a single experimental protocol, or whether data from different protocols were merged in the training set. Ideally, all (Q)SARs should be developed by using data generated by a single protocol, but this is rarely feasible in practice. In the case of commercially developed models, information on the training sets is not always made publicly available.
2. According to Principle 2, a (Q)SAR should be expressed in the form of an “unambiguous algorithm”. The intent of this principle is to ensure transparency in the description of the model algorithm. In the case of commercially developed models, a unambiguous definition of the algorithm(s) used is not always made publicly available.

### A defined endpoint

3. A (Q)SAR is a qualitative or quantitative relationship between chemical structure and the property or (biological) activity being modelled. The property or (biological) activity is called the “endpoint”, whereas the form of the relationship is called the “algorithm”.
4. The endpoint is often determined in accordance with an experimental protocol, and in the case of an endpoint of regulatory interest, with a test guideline (e.g. an EU Test Method or an OECD Test Guideline). Commonly used test guideline endpoints that are used in the assessment of chemicals are listed in Table 2.1. In the context of a test method, a well-defined endpoint is a property or effect that is measured according to a well-defined (standardized) set of experimental conditions.

5. (Q)SARs are often developed for well-defined (test method) endpoints. However, if the development of (Q)SARs is compared with that of test methods, several difference of emphasis can be noticed. For example, it is often assumed that the applicability domain (AD) of a test method is “global” in terms of its coverage of physicochemical space, whereas the AD of a (Q)SAR is generally assumed to be limited in one more respects. This means that, in general, a single (Q)SAR can only be expected to give comparable data to a test method when it is applied within its AD, and multiple (Q)SARs are often needed to make reliable predictions for a diverse range of chemicals. For this reason, research is being carried out into the development of methods and tools for combining the use of (Q)SARs.

6. Another difference between (Q)SARs and test methods is that the defined endpoints of many (Q)SARs are actually different to test method endpoints. In such cases, it should be considered whether the (Q)SAR gives “equivalent” data, which may be useful for hazard and risk assessment purposes, but it should not be assumed that the (Q)SAR gives “comparable” data, in the sense that the (Q)SAR data is directly predictive of the experimental data.

7. Even if a (Q)SAR is developed for a well-defined test method endpoint, the nature of the test method will affect the feasibility of producing a meaningful and predictive model. For example, it has been possible to develop reliable QSAR models for acute toxicity in fish, because the experimental protocol results in a steady-state concentration of the chemical between the blood and exposure medium. In such cases, it has been possible to relate quantitative variations in the endpoint (acute lethal toxicity) to (quantitative) variations in physicochemical structure. In contrast, it has been more difficult to develop meaningful and predictive models of acute oral toxicity in mammalian organisms, because it is difficult to separate the effect of chemical structure on potency from the effect of the organism on the kinetics of the chemical. Variations in the internal exposure conditions make it difficult to identify to inherent potency of the chemical.

### **Unambiguous algorithm**

8. The need for a (Q)SAR to be associated with an unambiguous algorithm reflects the need of the end-user to understand how the estimated value was generated, and to be able to reproduce the algorithm and/or the estimates. When the underlying algorithm of a (Q)SAR model is not transparent to the user, the model is sometimes referred to as a “black box”.

9. Principle 2 refers to the algorithm used to make a prediction of an endpoint on the basis of one or more descriptors. In many cases, it is possible to be transparent about this algorithm without necessarily being transparent about the (mathematical) method used to develop the algorithm. For example, a regression-based QSAR can be defined explicitly without any description of the regression approach. In addition, an expert rule can be stated explicitly, without any indication of how the rule was developed. In some cases, expert rules have been generated by automated procedures, so in principle the same procedures could be used to re-derive the rules. However, in other cases, expert rules simply codify the knowledge of experts, so no automated procedure can be used to re-derive the rule.

10. For certain types of model, the definition of the algorithm is more closely associated with the way in which it was derived (e.g. a neural network model which includes both a learning process and a prediction process).

11. An important component of an unambiguous algorithm is the availability of information on the descriptors used to link chemical structure with the predicted endpoint. Broadly, descriptors can be distinguished into three main types:

- a) descriptors based on chemical graph theory
- b) descriptors based on experimental measurements
- c) descriptors based on theoretical quantum mechanical calculations

Descriptors based on chemical graph theory are sometimes called topological descriptors. Examples include the molecular connectivity indices (10,11,12,13). Descriptors based on experimental measurements (e.g. logP) have traditionally been used most widely, as reviewed by Verhaar *et al.* (14). Increasingly, methods for predicting such descriptors are being used as a substitute for the experimental data. The third category of descriptors refers to a range of descriptors for predicting the electronic properties of molecules. These descriptors are either generated by semi-empirical models or more precise (and computationally intensive) *ab initio* methods. Some of the properties calculated by these methods can be measured (e.g. the HOMO energy) whereas others cannot be directly measured (e.g. LUMO energy). Some of these descriptors are useful for predicting chemical reactivity, and therefore for modelling chemical toxicity that results from reactions with cellular chemicals and macromolecules (14,15,16).

12. It is important to distinguish between the transparency of the algorithm and its mechanistic interpretability. For example, a statistically-based QSAR can be transparent in terms of its predictor variables and coefficients, but the descriptor variables themselves may not have an obvious physicochemical meaning or plausible causal link with the endpoint being modelled. The mechanistic interpretation of (Q)SARs is addressed in Chapter 5.

## Conclusions

13. For (Q)SAR purposes, a well-defined endpoint should ideally be based on experimental data generated by a standardised test protocol. In the case of (Q)SARs developed by using data from different protocols, the differences in the experimental conditions should not lead to significantly different values of the endpoint.

14. In the case of QSARs, i.e. quantitative relationships based on numerical measures of chemical structure, a well-defined endpoint is an endpoint that can be quantified in a way that reflects differences between chemical structures.

15. Transparency in the (Q)SAR algorithm can be provided by means of the following information:

- a) An explicit definition of the mathematical form of a QSAR model, or of the decision rule (e.g. in the case of a SAR)
- b) Definitions of all descriptors in the algorithm, and a description of their derivation
- c) Details of the training set used to develop the algorithm.

**Table 2.1 Endpoints associated with EU Test Methods and OECD test guidelines**

<b>Physicochemical Properties</b> Melting Point Boiling Point Vapour Pressure K <sub>ow</sub> octanol/water partition coefficient K <sub>oc</sub> organic carbon/water partition coefficient Water Solubility
<b>Environmental Fate</b> Biodegradation Hydrolysis in water Atmospheric Oxidation Bioaccumulation
<b>Ecological Effects</b> Acute Fish Long-term Toxicity Acute Daphnid Alga Terrestrial toxicity
<b>Human Health Effects</b> Acute Oral Acute Inhalation Acute Dermal Skin Irritation Eye Irritation Skin Sensitisation Repeated Dose Toxicity Genotoxicity ( <i>in vitro</i> , bacterial cells) Genotoxicity ( <i>in vitro</i> , mammalian cells) Genotoxicity ( <i>in vivo</i> ) Reproductive Toxicity Developmental Toxicity Carcinogenicity



# **CHAPTER 3**

## **(Q)SAR APPLICABILITY DOMAIN**

## CHAPTER 3: (Q)SAR APPLICABILITY DOMAIN

### Summary of Chapter 3

Chapter 3 provides guidance on how to interpret OECD validation principle 3 that “a (Q)SAR should be associated with a defined domain of applicability”. This principle expresses the need to establish the scope and limitations of a model based on the structural, physicochemical and response information in the model training set. The importance of the principle lies in the fact that a given model can only be expected to give reliable predictions for chemicals that are similar to those used to develop the model. Predictions that fall outside the applicability domain (AD) represent extrapolations, and are less likely to be reliable. When applying a (Q)SAR, it is important to know whether its AD is known, and whether it is being used inside or outside of this boundary. In its simplest form, the assessment of whether a chemical is located in the AD can be expressed categorically (i.e. yes or no). For a quantitative assessment, it is possible to associate a confidence interval with the AD, to determine the degree of similarity between the chemical of interest and the model training set. This chapter begins by explaining of the need for defining the AD (paras 1-4), before introducing some basic concepts and definitions (paras 5-11). The chapter then provides a review of different methods that are currently available or under development for identifying and quantifying the applicability domain, with some examples to illustrate their applicability (paras 12-34). It is emphasised that the subject of the (Q)SAR AD is an evolving field of research, and some research needs are presented in the concluding remarks of the chapter (para 36-40).



## Introduction

1. The principle that “a (Q)SAR should be associated with a defined domain of applicability” expresses the need to provide supporting information on the applicability of a (Q)SAR to unknown chemicals. This need is based on the fact that (Q)SARs are reductionist models which are inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions. In principle, every (Q)SAR model can be associated with an applicability domain (AD), even though this has not always been done explicitly in the QSAR literature.

2. Information on the AD helps the user of the model to judge whether the prediction for a new chemical is reliable or not. The definition of the AD is based on the assumption that a model is capable of making reliable predictions only within the structural, physicochemical and response space that is known from its training set. Thus, the model fit, robustness and predictivity determined by statistical methods (see Chapter 4) are meaningful only if they are used for chemicals in the AD. Even within the AD of a model, different degrees of confidence can be associated with different predictions.

3. The assessment of whether a chemical falls within the AD of a model is based on an assessment of the similarity between the chemical and the training set. Since there are many different ways of expressing similarity (often defined in physicochemical properties), it follows that many different methods for defining the AD can be developed. Indeed, a variety of methods have been proposed, each of which is associated with strengths and limitations (17).

4. The third OECD validation principle does not imply that each (Q)SAR is associated with a single AD. There is no absolute boundary between reliable and unreliable predictions. This can be fixed by the model user, according to the performance characteristics of the model and the context in which the model is being applied. In general, there is a trade-off between the breadth of the AD and the overall reliability of prediction: the broader the AD, the fewer chemicals that can be predicted with a given reliability.

## Basic Terms and Concepts

5. Several definitions of the AD can be found in the (Q)SAR literature, but probably the most broadly applicable definition is the following (17):

“The applicability domain of a (Q)SAR model is the response and chemical structure space in which the model makes predictions with a given reliability. ”

6. In this definition, chemical structure can be expressed by information on physicochemical properties and/or structural fragments, and the response can be any physicochemical, biological or environmental effect that is being predicted (i.e. the defined endpoint, see Chapter 2). The relationship between chemical structure and the response can be expressed by a variety of SARs and QSARs.

7. The AD concept should be applied in a model-specific manner. Thus, every model should be associated with its own AD. It depends not only on the chemicals in the training set but also on the descriptors and (statistical) approach used to develop the model. Thus, the same training set could in principle be used to develop multiple QSARs, differing in terms of their descriptors and/or mathematical form, and with different ADs.

8. Ideally, the AD should be defined and documented by the model developer. However, this information is often lacking in the reports of (Q)SAR studies. In principle, it should be possible for an independent (Q)SAR practitioner to define (or confirm) the AD of an existing model, provided that a sufficient amount of background information is available. This information should include: a statement of the unambiguous model algorithm (see Chapter 2), details of the training set (chemical identification, descriptors and endpoint values), and details of the (statistical) method to derive the model.

9. Ideally, the AD should express the structural, physicochemical and response space of the model. This is because the best assurance that a chemical is predicted reliably is to have confirmation that the chemical is not an outlier in terms of its structural fragments (structural domain), its descriptor values (physicochemical domain) or its response values (response domain). In some cases, a model can predict reliably beyond its physicochemical domain, especially if it is still within its structural domain.

10. Even though a well-defined AD helps the user of the model to assess the reliability of predictions made by the model, it should not automatically be assumed that all predictions within the defined AD are necessarily reliable. In practice, a prediction could still be unreliable even though the chemical lies within the established structural and physicochemical domains of the model. This could occur in cases where the chemical of interest acts by a different mechanism of action, not captured by the model. If more than one such chemical is discovered, the QSAR practitioner could either try to refine the model, to accurately predict the outliers, or could try to define an exclusion rule. The need to account for such outliers has also led to the concept of the mechanistic domain. Thus, for some models, the application of OECD validation principle 3 is linked with the application of principle 5 on mechanistic interpretation.

11. The concept of AD should be applied only to statistically validated (Q)SARs, i.e. (Q)SARs that are based on statistically significant relationships, and not on random models, where fundamental statistical principles are violated (see Chapter 4).

## **Recommendations for Practitioners**

12. Historically, the first QSAR models were developed for homologous series of chemicals. Although these models may have limited use today, they are helpful to illustrate how the concept for the AD can be applied. For example, if one knows the narcotic effects of the primary alcohols ethanol, propanol, butanol, hexanol and heptanol, then one can predict the narcotic effect of pentanol by the linear relationship between the narcotic effect and molecular weight (MW). Pentanol is in the AD of this simple model because it is a structural homologue of the other alcohols and has a MW intermediate to two other alcohols. The alcohols methanol and the octanol, however, would not be considered in the AD of the model because, because while they are structural homologues of the other alcohols, they have MW values lower than ethanol and greater than heptanol, respectively (Figure 3.1).

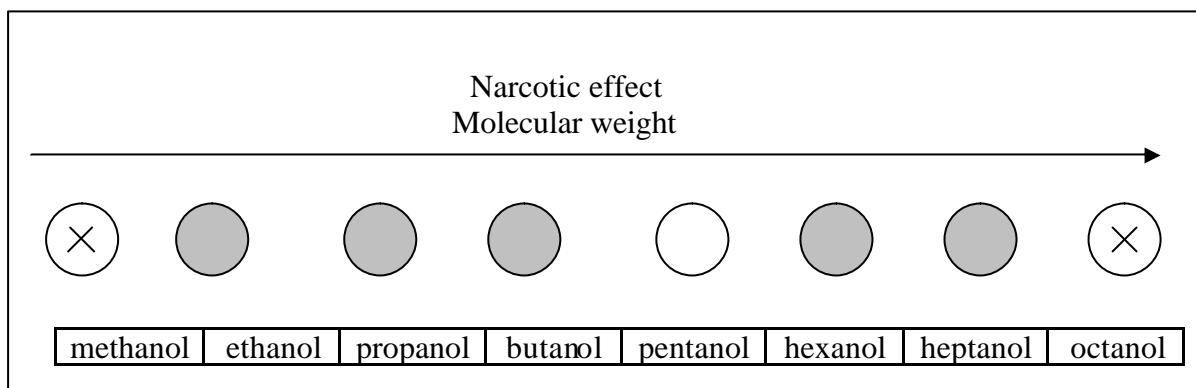


Figure 3.1

13. Other examples support the same reasoning. For example, Hermens *et al.* (18) have shown that increasing the number of carbon atoms in a homologue series of aldehydes above 10 leads to a change of the mechanism of action. The consequence is that the relationship between the toxicity and the octanol-water partition coefficient ( $\log K_{ow}$ ), found for lower members of the series, does not hold true for the higher members. In another example, Schultz *et al.* (19) showed that acrolein, the first chemical in the series of  $\alpha,\beta$ -unsaturated aldehydes, was considerably more toxic than predicted by the relationship between  $\log K_{ow}$  and toxicity for the other  $\alpha,\beta$ -unsaturated aldehydes.

14. In addition to the physicochemical and structural domains, an additional useful element in the AD definition is an understanding of the mechanism of action (MOA) of the chemicals used to develop a model (i.e. the mechanistic domain). For example, the phenols and the anilines (if not complicated by more reactive moieties) demonstrate polar narcosis in aquatic organisms (20) even though they belong to different chemical classes. Thus, the effects of chemicals belonging to both chemical classes can be predicted by a single model provided the chemical does not go beyond the range of physicochemical parameters used to develop the model. The grouping of chemical classes into single QSARs is endpoint-specific because the different classes might not behave in the same way for a different endpoint (e.g. mutagenicity). In fact, aromatic amines have considerable potential to cause mutations whereas phenols do not.

15. Chemicals that contain multiple functional groups deserve special attention. Such chemicals might exhibit enhanced effects as a result of synergism or even exhibit a different MOA. Such chemicals are likely to be outliers to well established relationships. An example is provided by the  $\alpha$ -halogenated esters (21), in which the presence of a halogen atom on an aliphatic hydrocarbon chain does not alter the narcosis MOA for aquatic toxicity. Aliphatic esters also act as narcotics in aquatic organisms. However, the presence of a halogen atom at the  $\alpha$ -position to the carbonyl group of an aliphatic ester results in a drastic increase of toxicity due to the fact that this arrangement of atoms undergoes an  $SN_2$  reaction (the halogen atom being the leaving group) with macromolecules.

16. The identification of special atom arrangements (toxicophores) that cause certain types of toxicity provides a way of defining mechanistic domains. Expert judgement is required since the expected toxicological profile could be modulated by the presence of additional functional groups (modulators), which may increase or decrease the toxicity. For example,

the methyl groups usually increase the toxicity due to increased lipophilicity without changing the MOA. Thus, the methylphenols are slightly more toxic to fish than the parent phenol (22). However, methyl groups can also block completely the toxicophore; for example the methyl groups in the *tert*-butyl group decrease the toxicity of *tert*-butyl acrylate (23). The presence of a bulky substituent next to a reactive group is one reason why a chemical might fall outside the expected mechanistic domain. The properties of such chemicals or are usually overestimated.

17. Inaccuracy of prediction can appear also if a chemical undergoes metabolic transformation. Such chemicals appear outliers from many different (Q)SAR models irrespectively of whether the model was developed on a mechanistic basis or statistically. The reason for miss-prediction in this case is that the chemical that causes the effect is different from the chemical that was introduced to the biologic system and these out-of-the-domain chemicals are usually most difficult to identify *a priori*. An example could be given with 1,2- and 1,4-dihydroxybenzenes that exhibit enhanced toxicity because of transformation to 1,2- and 1,4-quinones with strong electrophilic potential, or formation of free-radical species (24).

18. At present, the identification of mechanistic domains relies heavily on expert judgement. There are, however, some software tools that can assist in the identification of potential toxicophores and modulators. An example is the DEREK software (Lhasa Ltd), an expert system that applies knowledge-based rules for toxicity prediction. A similar functionality is available in HazardExpert (Compudrug.Inc.), which issues an alert if a toxic fragment is found in the query molecule. Another program for toxicity prediction, MULTICASE (Multicase Inc.), evaluates the structural features of molecules from non-congeneric series and identifies substructural fragments that are considered responsible for a certain type of activity. The TOPKAT software (Accelrys Inc.) uses an initial classification into chemical classes before applying quantitative models for toxicity prediction. Various software products incorporate knowledge about metabolism and can therefore be used to anticipate the metabolites of the chemical of interest. These systems include CATABOL (Laboratory of Mathematical Chemistry, Bulgaria) and MetabolExpert (CompuDrug Inc.).

19. As expressed by OECD principle 5, it is useful if the (Q)SAR can be interpreted in mechanistic terms (see Chapter 5). However, this is not always possible, in part because the underlying mechanisms of many toxic effects are simply not known. It is also not essential, since robust models can be develop by purely statistical means. For the purposes of this chapter, it should be noted that the different assumptions behind mechanistic and statistical (Q)SARs imply the need for different types of tool for defining the AD.

20. If a (Q)SAR is based on physicochemical descriptors, the interpolation space (i.e. its coverage), defined by its descriptors, should be characterised. The interpolation space of a one-descriptor model is simply the range between the minimum and the maximum value of that descriptor, as observed in the training set of the model. The interpolation space of multi-descriptor models is more complex. Several statistical methods can be applied to characterise the interpolation space, as described below.

21. The simplest method for describing the AD is to consider the ranges of the individual descriptors. This approach is based on the assumption that the descriptor values follow a normal distribution, and could therefore be unreliable if this assumption is violated. A limitation of this approach is that the AD may include internal empty spaces, i.e. interpolation regions where the relationship is not proved (Figure 3.2). Another possible limitation is the

fact that intercorrelation between the descriptors is not taken into account, unless the individual descriptors are replaced by their principal components.

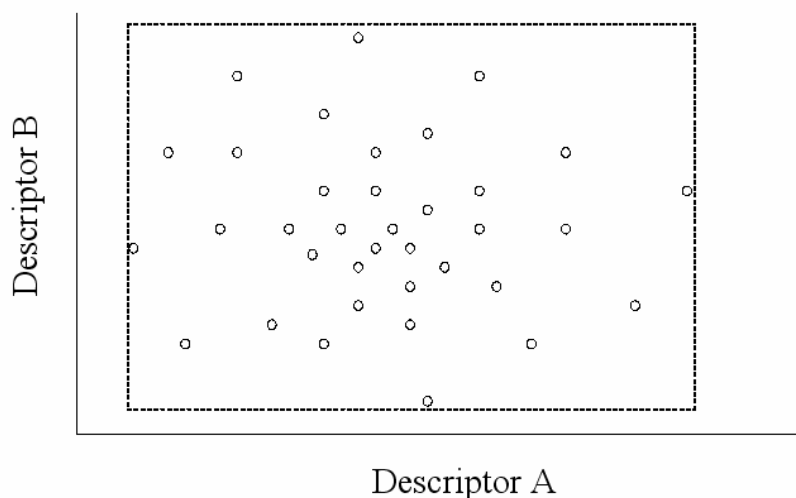


Figure 3.2

22. A more advanced method for defining the interpolation space of a model is to define the smallest convex area that contains the descriptors of the training set. However, this method does not solve completely the problem of empty spaces between the chemicals of the training set. In addition, for models that contain many descriptors, the calculation of the convex area becomes a time-consuming computational problem (see Figure 3.3).

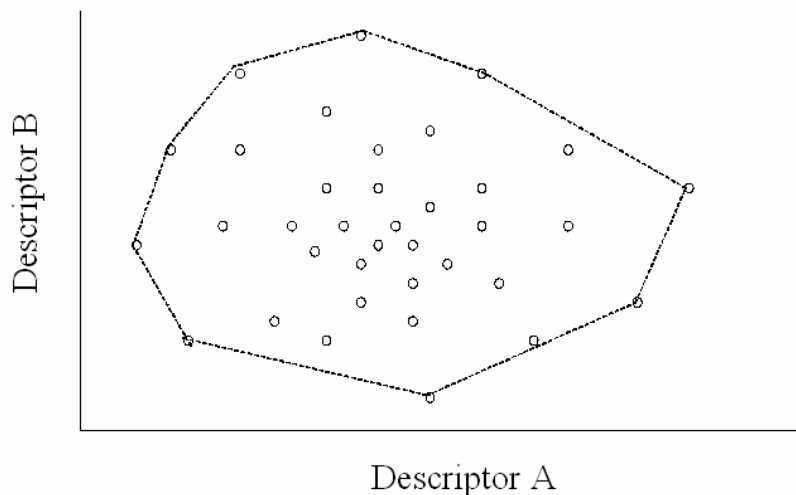


Figure 3.3

23. A different approach to defining the AD is based on a calculation of the distance between a query chemical and a defined point in the descriptor space of the model (typically, the centroid of the training set). A detailed review of methods is given by Jaworska *et al.* (25). Different methods following this approach can be applied (e.g. Euclidean, Mahalanobis, Manhattan distance). The advantage of the distance (also called geometric) approach is that

confidence levels can be associated with the AD by drawing iso-distance contours in the interpolation space. The disadvantage is again the assumption of a normal distribution for the underlying data. This means that the contours are drawn solely on the basis of the distance from the centroid, and the population of the regions between two iso-distance contours is not taken into account.

24. A common approach to distance analysis is to use the Hotelling's test and the associated leverage statistics. The leverage of a chemical provides a measure of the distance of the chemical from the centroid of its training set. Chemicals in the training set have leverage values between 0 and 1. A "warning leverage" ( $h^*$ ) is generally fixed at  $3p/n$ , where  $n$  is the number of training chemicals, and  $p$  the number of descriptors plus one. A leverage value greater than the warning leverage is considered large.

25. The leverage is a useful statistic in both QSAR development and application. During QSAR development, chemicals with high leverage unduly influence the regression parameters of the model, and yet do not appear as statistical outliers (the regression line is forced near the high leverage chemical). It may therefore be appropriate to exclude such chemicals from the training set. During the application of a QSAR, chemicals with high leverage are likely to be outside the descriptor space of the model, and therefore the predictions for such chemicals could be unreliable. The leverage approach is illustrated in Gramatica *et al.* (26) and Pavan *et al.* (27).

26. As with all statistical methods based on physicochemical descriptors, the leverage approach needs to be applied with care. While the observation that a chemical has a large leverage indicates that it is outside the descriptor coverage of the model, a chemical with small leverage can also be outside the AD for other reasons (e.g. a presence of a toxicophore that is not present in the training set). The inability to discriminate unequivocally between chemicals that are inside and outside the AD is common to all statistical methods based on physicochemical descriptors, and this should be taken into account when applying the concept of the AD.

27. To visualise the outliers in a model, i.e. outliers in both the descriptor space and the response space, a plot of standardised residuals ( $R$ ) *vs* leverages (or hat values,  $h$ ), called the Williams graph is sometimes used. An illustration of the Williams plot, taken from Pavan *et al.* (27), is given in Figure 3.4a. This shows the training set of 86 chemicals for a polar narcosis model of acute toxicity to the fathead minnow (Verhaar *et al.*, 20) as well as a test set of 8 chemicals for which the model was used to make predictions. It can be seen that 6 chemicals in the training set have leverages greater than the warning leverage (0.07), as do 2 of the test chemicals. The corresponding regression line for the model is shown in Figure 3.4b.

28. The most advanced statistical methods that are currently applied for identifying the (Q)SAR AD are probability density distribution-based methods. The probability density function of a data set can be estimated by parametric and non-parametric methods. The parametric methods assume a standard distribution (e.g. Gaussian or Poisson distribution) while the non-parametric methods (e.g. kernel density estimation function) make no assumptions about the data distribution. An advantage of non-parametric methods is the ability to identify internal empty spaces and, if necessary, to generate concave regions around the borders of the interpolation space to reflect the actual data distribution. It has been argued that the probability density approach is more robust than the range, distance and leverage

approaches (28). However, it is also more restrictive in terms of the chemical space that falls in the AD of a model.

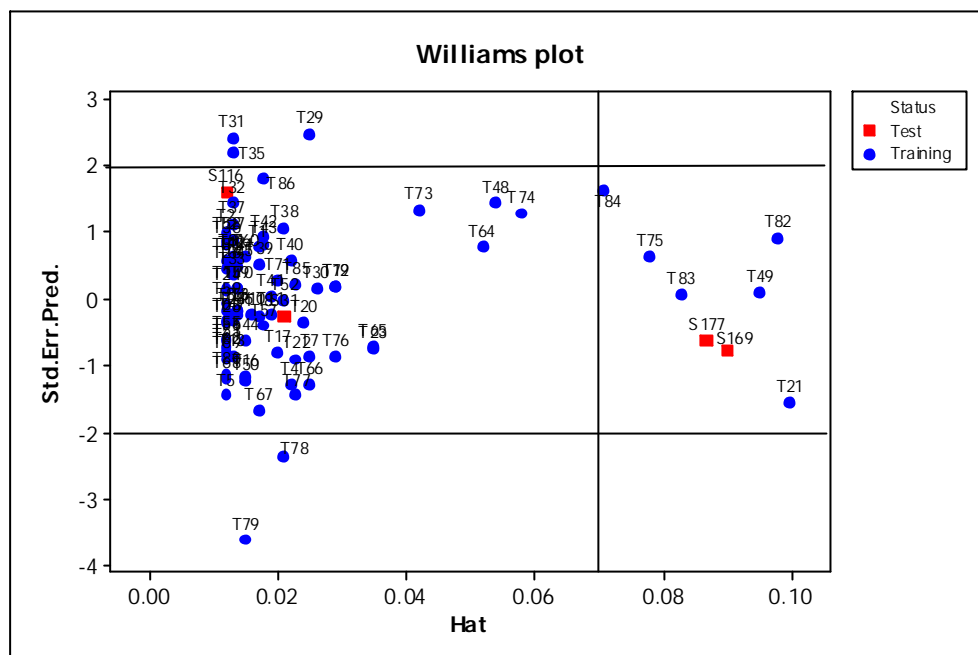


Figure 3.4a

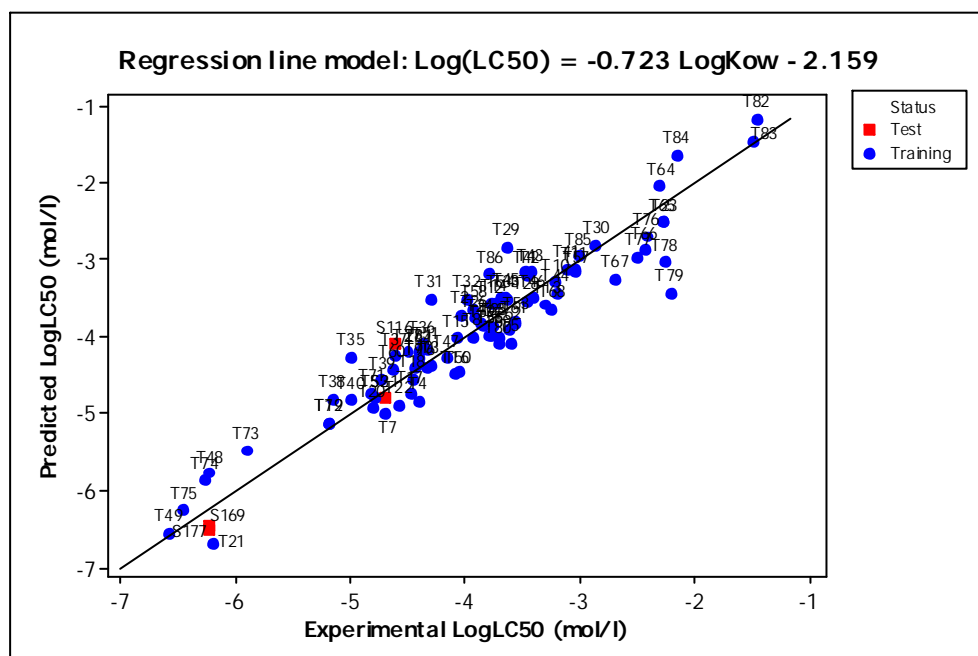


Figure 3.4b

29. While some of the described statistical methods are available in a standard statistical packages (e.g. MINITAB, STATISTICA, SYSTAT), they are not adapted to meet the needs of (Q)SAR developers and users. In contrast, a user-friendly software package called Ambit

Disclosure being developed under the auspices of CEFIC LRI can be used to calculate the interpolation space by knowing the values of the dependent (endpoint) and independent (descriptors) variables used in a given model. The AD methods incorporated in the software are independent of the modelling technique and require only transparency of the training set. A free download is available on the internet (29).

30. Ideally, the coverage of the training set should be accompanied by information on the structural or physicochemical similarity between the query molecule and the (Q)SAR training set. The similarity can be expressed in a qualitative or quantitative manner. Preferably, some mechanistic rationale should be given of whether the query chemical represents a mechanism common to a group of chemicals in the (Q)SAR training set. However, when such an assessment is not possible, a statistical expression of similarity can be obtained.

31. One possible approach is to split the query molecule in fragments and to check whether all the fragments are represented in the training set of the model. The higher the occurrence of the query fragments in the training set, the higher the confidence that the query chemical of interest can be predicted reliably. This approach is adopted in the MultiCASE software and in the Leadscape platform (Leadscape Inc.). These programs issue a warning message that a chemical is outside of the AD of the model if encounters an unknown fragment.

32. A quantitative expression of similarity can be obtained by calculating the Tanimoto coefficient. The Tanimoto coefficient is the ratio of shared substructures to the number of all substructures that appear in the reference chemical in the training set. The Tanimoto coefficient varies between 0 (total lack of similarity) to 1 (the query chemical has an identical constitution to the reference chemical). It is important to remember that the Tanimoto coefficient does not provide a unique measure of similarity - its meaning is based on how structural fragments are defined for the purposes of the comparison. Thus, two chemicals that are similar with a Tanimoto coefficient of 0.8 on the basis of one set of fragments may not be similar when compared by using a different set of fragments. Algorithms for calculating Tanimoto similarity coefficients are incorporated in several software products, including Ambit disclosure software and the Leadscape platform.

33. A different approach needs to be adopted if multiple (Q)SAR models are being used for the prediction of the same endpoint. Theoretically, different models have different ADs. If a query chemical falls within the intersection of the ADs of the different models, the confidence of the overall prediction after combining (by averaging or other transformation) the individual predictions should be greater than the confidence associated with the prediction of a single model. However, it is expected that the common AD will be narrower for multiple models, thus restricting the number of potential chemicals that could be predicted. An example of the use of multiple models is provided by Tong *et al.* (30), who used a decision forest (i.e. multiple comparable and heterogeneous decision trees).

34. Recently, a stepwise approach for determining the model AD has been proposed by Dimitrov *et al.* (31). It consists of four stages. The first stage identifies whether a chemical falls in the range of variation of physicochemical properties of the model. The second step defines the structural similarity between the query chemical and chemicals correctly predicted by the model. The third stage comprises a mechanistic check by assessing whether the chemical contains the specific reactive groups hypothesised to cause the effect. The fourth and final stage is a metabolic check, based on an assessment of the probability that the



chemical undergoes metabolic activation. The four stages are applied in a sequential manner. The advantage of processing query chemicals through all four stages is the increased reliability of prediction for those chemicals that satisfy to all four conditions for inclusion in the AD. The cost of applying this rigorous, multiple AD approach is that the number of chemicals for which reliable predictions are eventually made is reduced compared to the use of a single AD method.

### ***Comparing applicability domains with the spaces of regulatory inventories***

35. Defining the AD of a model not only provides a means of increasing the confidence associated with predictions inside the domain, but also of assessing the applicability of the model to a given regulatory inventory of chemicals. A model that gives highly accurate predictions for narrow chemical classes that are not covered by the regulatory inventory of interest would be of questionable value. A number of investigations have addressed the need to screen and prioritise chemical inventories established under different legislations in OECD Member Countries (32, 33, 34). Among the most commonly screened regulatory inventories are those of the High Production Volume Chemicals, Existing Substances, and inventories of pesticides and biocides. Less information is publically available regarding the inventories of New Substances, mainly because of confidentiality considerations. In addition, these inventories are periodically updated with new chemicals, which implies the need for iterative development of (Q)SAR models (35, 36) to expand their domains and adapt them to the regulatory domains of concern. An approach for comparing the AD with a regulatory domain is illustrated in a study (37) in which the AD of a QSAR for estrogenic potential is compared with the domain of the EINECS inventory (the list of Existing Substances in the EU). In this study, the physicochemical space of the EINECS inventory is characterised by using the descriptors in the QSAR model.

### **Concluding remarks**

36. The third OECD principle on the need for a defined AD should be considered in combination with the fourth OECD principle on the need to characterise the statistical validity of a model, since an understanding of the AD can increase or decrease the confidence in a given (Q)SAR estimate. It should be noted, however, that the use of AD methods will never provide absolute certainty in the (Q)SAR estimates: a query chemical may appear to be within the defined AD, and yet the prediction could still be unreliable; conversely, the query chemical may appear to be outside the defined AD, and yet the prediction could be reliable.

37. The model user should therefore be aware that AD methods, like other (statistical) methods discussed in this Guidance Document, provide a useful means of supporting decisions based on the additional use of expert judgement, but they cannot in themselves make the decisions.

38. Numerous AD methods have been proposed based on the following considerations: structural features, physicochemical descriptor values, response values, mechanistic understanding, and metabolism. On this basis, it is useful to conceptualise the AD of a model as the combination of one or more elements relating to the structural, physicochemical, response, mechanistic and metabolic domains. While these different types of domains provide useful distinctions, they should not be assumed to be mutually exclusive. For example, the

structural fragments present in a molecule will affect its physicochemical descriptors, its response value, and its mechanistic behaviour.

39. The different AD methods should not be seen as in competition with one another, since the combined use of multiple AD methods should give a higher assurance that query chemicals are predicted accurately by a (Q)SAR model. Inevitably, there is a trade-off between the breadth of applicability of a model and the reliability of the predictions within the domain: the broader the scope of the model, the lower the overall reliability of prediction. The user of a model therefore needs to strike an appropriate balance between the level of confidence in the predictions resulting from AD considerations and the number of reliable predictions that are determined.

40. Attempts to formalise and quantify the concept of the AD are relatively recent, which means that it is still a difficult concept to apply in regulatory practice. Thus, a considerable amount of research and development is still needed to further develop AD methods, as well as an understanding of the applicability of these methods. For example, the following research needs can be identified:

- a) the development of confidence limits associated with the AD;
- b) the development of AD methods for structural alerts and fragment-based QSAR methods;
- c) the assessment of AD methods with a view to better understand their strengths, limitations and applicabilities;
- d) the development of automated tools that facilitate the application of AD methods in an integrated manner with traditional statistical methods.

# **CHAPTER 4**

## **STATISTICAL VALIDATION**

## CHAPTER 4: STATISTICAL VALIDATION

### Summary of chapter 4

This chapter provides guidance on how to interpret OECD validation principle 4 that “a (Q)SAR should be associated with appropriate measures of goodness-of-fit, robustness and predictivity”. This principle expresses the need to perform statistical validation to establish the performance of the model, which consists of internal model performance (goodness-of-fit and robustness) and external model performance (predictivity), taking into account any knowledge about the applicability domain of the model (Chapter 3). The chapter starts with a brief introduction to principle 4 and statistical validation (1-4), followed by an explanation of some key terms and concepts (5-10). In paras 11-76, commonly used techniques for model development are then described and illustrated (multiple linear regression, partial least squares, classification modelling, neural network modeling) along with well-established statistical validation techniques for assessing goodness-of-fit, robustness and predictivity (cross-validation, bootstrapping, response randomisation test, training/test splitting, external validation). In the context of these different techniques, the statistics that are commonly used to describe model performance are explained.

### Introduction

1. The regulatory acceptance and use of (Q)SAR models depends partly on what is known about their statistical performance. The need for information on model performance is expressed by OECD validation principle 4, according to which models should be associated with appropriate measures of goodness-of-fit and robustness (internal performance) and predictivity (external performance). The assessment of model performance is sometimes called statistical validation.
2. Statistical validation techniques are used during (Q)SAR development to find a suitable balance between two extremes: overfitted and underfitted models. The optimal model complexity is a trade-off between models that are “too simple” and therefore lacking in useful information and models that are too “complex” and therefore modelling noise (38,39). Statistical validation techniques provide various “fitness” functions that can be used by the QSAR practitioner to compare the quality of different models, and to avoid models that are too simplistic or too complex.
3. Statistical validation techniques also provide a means of identifying “spurious” models based on chance correlations, i.e. situations in which an apparent relationship is established between the predictor and response variables, but which is not meaningful and not predictive (40,41,42).
4. The statistical validation techniques described in this chapter should be considered in combination with any knowledge about the applicability domain (AD) of the model, since the choice of chemicals during model development and validation affects the assessment of performance. In particular, chemicals that are outside the AD during model development may unduly influence the regression parameters of the model, thereby affecting its robustness.

Chemicals that are outside the AD during model validation are unlikely to be predicted with the desired level of reliability.

## Basic concepts

5. This section provides an explanation of some key concepts that are needed to understand the remainder of the chapter.

6. The QSAR modelling process starts with the compilation of a data set which is often divided into a *training set*, used to derive the model through the application of a statistical method, and a *test set*, containing chemicals not used in the derivation of the model. The variables in the model are chosen to optimise model complexity, and are referred to as *predictors*. In QSAR modelling, the predictors are (molecular) descriptors.

7. The model derived from the training set is used to predict of the response data in both the training and the test sets. The *accuracy* of prediction for a given chemical is the closeness of an estimate/prediction to a reference value. The greater the proportion of accurate predictions, the more *reliable* the model.

8. Predictions for chemicals in the training set are used to assess the *goodness-of-fit* of the model, which is a measure of how well the model accounts for the variance of the response in the training set. The generation of predictions within the range of predictor values in the training set is called *interpolation*, whereas *extrapolation* is the generation of a prediction outside the range of values of the predictor in the sample used to generate the model. The more removed the predicted value from the range of values used to fit the model, the more unreliable the prediction becomes, because it is not certain whether the model continues to hold.

9. The *robustness* of model refers to the stability of its parameters (predictor coefficients) and consequently the stability of its predictions when a perturbation (deletion of one or more chemicals) is applied to the training set, and the model is regenerated from the “perturbed” training set.

10. Predictions for chemicals in the test set are used to assess the *predictive ability* of the model, which is a measure of how well the model can predict of new data, which not used in model development. In this document, predictive ability is used synonymously with *predictive capacity*, *predictive power* and *predictivity*.

## Recommendations for practitioners

### *Multiple Linear Regression (MLR)*

11. Multiple linear regression (MLR) is the traditional statistical approach for deriving QSAR models. It relates the dependent variable  $y$  (biological activity) to a number of independent (predictor) variables  $x_i$  (molecular descriptors) by using linear equations (Eq. 1, Table 4.1).

12. **Estimating the regression coefficients.** Regression coefficients  $b_j$  in MLR model can be estimated using the least squares procedure by minimizing the sum of the squared residuals. The aim of this procedure is to give the smallest possible sum of squared differences between the true dependent variable values and the values calculated by the regression model.

13. **Assessing the relative importance of descriptors.** If the variables are standardized to have mean of zero and standard deviation of one, then the regression coefficients in the model are called *beta* coefficients. The advantage of *beta* coefficients (as compared to regression coefficients that are not standardised) is that the magnitude of these *beta* coefficients allows the comparison of the relative contribution of each independent variable in the prediction of the dependent variable. Thus, independent variables with higher absolute value of their *beta* coefficients explain greater part from the variance of the dependent variable.

### ***Measures of goodness-of-fit in Multiple Linear Regression***

14. **Assessing goodness-of-fit.** To assess goodness-of-fit, the coefficient of multiple determination ( $R^2$ ) is used (Eq. 2, Table 4.1).  $R^2$  estimates the proportion of the variation of  $y$  that is explained by the regression (43). If there is no linear relationship between the dependent and the independent variables  $R^2 = 0$ ; if there is a perfect fit  $R^2 = 1$ .  $R^2$  value higher than 0.5 means that the explained variance by the model is higher than the unexplained one. The end-user(s) of a QSAR model should decide what value of  $R^2$  is sufficient for the specific application of the model. One author has recommended that  $R = 0.9$  for in vitro data and  $R = 0.8$  for in vivo data can be regarded as good (44).

15. **Avoiding overfitting.** The value of  $R^2$  can generally be increased by adding additional predictor variables to the model, even if the added variable does not contribute to reduce the unexplained variance of the dependent variable. It follows that  $R^2$  should be used with caution. This could be avoided by using another statistical parameter – the so-called adjusted  $R^2$  ( $R^2_{adj}$ ) (Eq. 3, Table 4.1).  $R^2_{adj}$  is interpreted similarly to the  $R^2$  value except it takes into consideration the number of degrees of freedom. It is adjusted by dividing the residual sum of squares and total sum of squares by their respective degrees of freedom. The value of  $R^2_{adj}$  decreases if an added variable to the equation does not reduce the unexplained variance.

16. From the calculated and observed dependent variable values the standard error of estimate  $s$  could be obtained (Eq. 4, Table 4.1). The standard error of estimate measures the dispersion of the observed values about the regression line. The smaller the value of  $s$  means higher reliability of the prediction. However it is not recommended to have standard error of estimate smaller than the experimental error of the biological data, because it is an indication for an overfitted model (44).

17. The statistical significance of the regression model can be assessed by means of  $F$ -value (Eq. 5, Table 4.1). The  $F$ -value is the ratio between explained and unexplained variance for a given number of degrees of freedom. The higher the  $F$ -value the greater the probability is that the equation is significant. The regression equation is considered to be statistically significant if the observed  $F$ -value is greater than a tabulated value for the chosen level of significance (typically, the 95% level) and the corresponding degrees of freedom of  $F$ . The degrees of freedom of  $F$ -value are equal to  $p$  and  $n-p-1$ . Significance of the equation at the

95% level means that there is only a 5% probability that the dependence found is obtained due to chance correlations between the variables.

18. The statistical significance of the regression coefficients can be obtained from a *t*-test (Eq. 6, Table 4.1). It is used to test the hypothesis that the regression coefficient is zero. If the hypothesis is true, then the predictor variable does not contribute to explain the dependent variable. Higher *t*-values of a regression coefficient correspond to a greater statistical significance. The statistical significance of a regression coefficient using its *t*-value is determined again from tables for a given level of significance (similar to the use of *F*-value). The degrees of freedom of *t* are equal to  $n-p-1$  (corresponding to the degrees of freedom of the residual mean square). Statistical significance at the 95% level means there is only a 5% probability that the regression coefficient of a given variable is not significantly different from zero. The *t*-values are used to calculate the confidence intervals for the true regression parameters. These confidence intervals can also be used to check the significance of the corresponding regression coefficients. In practice the confidence intervals should be smaller than the absolute values of the regression coefficients in order to have statistically significant independent variables (**Error! Reference source not found.**).

### ***Partial Least Squares regression (PLS)***

19. Partial Least Squares (PLS), introduced by Wold *et al.* (45, 46), is a MLR method that allows relationships to be sought between an **X**-block of *p* predictors and a single **y** response (PLS1) or a **Y**-block of *r* responses (PLS2). Thus several activity variables, **Y**, i.e. profiles of activity, can be modelled simultaneously. An advantage of PLS is that it tolerates a certain amount of missing data. For instance, in the case of data set containing 20 compounds, 10-20% missing data can be tolerated (47).

20. **Information provided by PLS.** The purpose of PLS is to find a small number of relevant factors (**A**) that are predictive of **Y** and utilize **X** efficiently (48). The PLS model is expressed by a matrix of scores (**T**) that summarizes the **X** variables, a matrix of scores (**U**) that summarizes the **Y** variables, a matrix of weights (**W**) expressing the correlation between **X** and **U**(**Y**), a matrix of weights (**C**) expressing the correlation between **Y** and **T**(**X**), and a matrix of residuals (the part of data that are not explained by the model). For the interpretation of the PLS model a number of informative parameters can be used. The scores **t** and **u** contain information about the compounds and their similarities/dissimilarities with respect to the given problem. The weights **w** and **c** provide information about how the variables can be combined to form a quantitative relation between **X** and **Y**. Hence they are essential for understanding which **X** variables are important and which **X** variables provide the same information. The residuals are of diagnostic interest – large residuals of **Y** indicate that the model is poor and the outliers should be identified (47). PLS regression coefficients can be obtained after re-expression of the PLS solution as a regression model. When **X** values are scaled and centered and **Y** values are scaled, the resulting coefficients are useful for interpreting the influence of the variables **X** on **Y** (49, 50).

### ***Measures of goodness-of-fit in Partial Least Squares regression***

21. The quantitative measure of the goodness of fit is given by the parameter  $R^2$  (= the explained variation) analogous to MLR. PLS model is characterized by the following  $R^2$  parameters:

- $R^2(Y)$  – cumulative sum of squares of all dependent values (Y) explained by all extracted components
- $R^2(X)$  – cumulative sum of squares of all descriptor values (X) explained by all extracted components
- $R^2(Y)_{adj}$ ,  $R^2(X)_{adj}$  – cumulative  $R^2(Y)$  and  $R^2(X)$  respectively adjusted for the degrees of freedom

22. **Avoiding overfitting.** Depending on the number of components, near perfect correlations are often obtained in PLS analysis, due to the usually large number of included **X** variables. Therefore, the high  $R^2(Y)$  is not a sufficient criterion for the validity of a PLS model. A cross-validation procedure must be used and  $Q^2(Y)$  parameter must be calculated to select the model having the highest predictive ability (44). In contrast to  $R^2(Y)$ ,  $Q^2(Y)$  does not increase after a certain degree of model complexity. Hence, there is a zone, where there is a balance between predictive power and reasonable fit (48). According to the proposed reference criteria the difference between  $R^2(Y)$  and  $Q^2(Y)$  should not exceed 0.3. A substantially larger difference is indication for an overfitted model, presence of irrelevant **X**-values or outliers in the data (51).

23. **Identification of outliers.** As a measure of the statistical fit of the PLS model also the residual standard deviation (RSD) can be used, which corresponds to the standard deviation in the MLR. It should be similar in size to the known or expected noise in the system under investigation. The RSD can be calculated for the responses and predictor variables. The RSD of an **X** variable is indication for its relevance to the PLS model. The RSD of a **Y** variable is a measure of how well this response is explained by the PLS model. The RSD of an observation in the **X** or **Y** space is proportional to the observation distance in the hyper plane of the PLS-model in the corresponding space (DModX and DModY). The last ones give an information about the outliers in **X**- and **Y**-data (48, 50).

### ***Classification Models (CMs)***

24. Chemicals are sometimes classified into two (e.g. active/inactive) or more pre-defined categories, for scientific or regulatory purposes. For scientific purposes, the biological variability of certain endpoints is sometimes too large to enable reasonable quantitative predictions, so that the data is converted into one or more categories of toxic effect. Otherwise, in regulatory toxicology, binary classification systems are commonly used to provide a convenient means of labelling chemicals, according to their hazard.

25. Classification-based QSARs, also referred as classification models (CMs), can be developed using a variety of statistical methods. Among the methods appropriate for the development of linear CMs, multivariate discriminant analysis (MDA), logistic regression (LR), and decision or classification trees (CT), among others, have been extensively described in the literature (52). Also, rule-based models expressed in symbolic “if... then” decision rules, can be derived from the CMs. For the models associated with non-linear boundaries, embedded cluster modelling (ECM) (53), neural networks (NN), and k-nearest neighbour (k-NN) classifiers can be used.



## Measures of goodness-of-fit in classification models

26. The goodness-of-fit of a CM can be assessed in terms of its Cooper statistics, which were introduced in the late seventies to describe the validity of carcinogen screening tests (54**Error! Reference source not found.**). Cooper statistics, based on a Bayesian approach (55, 56) has been extensively applied to assess the results of classification (Q)SAR models (51, 57). Bayesian-based methods can also be used to combine results from different cases, so that judgments are rarely based only on the results of a single study but they rather synthesize evidence from multiple sources. These methods can be developed in an iterative manner, so that they allow successive updating of battery interpretation.

27. In a CM, the results of the classification can be arranged in the so-called *confusion* or *contingency matrix* (58), where the rows represent the reference classes ( $A_g$ ), while the columns represent the predicted classes assigned by the CM ( $A_g'$ ). Table 4.2 illustrates the general form of a contingency matrix for the general case of  $G$  classes.

28. **Interpreting the contingency matrix.** The main diagonal ( $c_{gg'}$ ) represents the cases where the true class coincides with the assigned class, that is, the number of objects correctly classified in each class, while the non-diagonal cells represent the misclassifications. Overpredictions are to the right and above the diagonal, whereas underpredictions are to the left and below the diagonal. The right-hand column reports the number of objects belonging to each class ( $n_g$ ), whereas the bottom row reports the total number of objects assigned to each class according to the CM ( $n_{g'}$ ).

29. **Setting the importance of misclassifications.** Depending on the intended use of the CM, some classification errors may be considered “worse” than others. In order to quantify such error, the *loss matrix* ( $L$ ), which has the same structure as the contingency matrix, can be used (Table 4.3). It can be considered as a matrix of weights for the different types of classification errors, where the non-diagonal elements quantify the type of error in the classification.

30. According to this matrix of weights, the classification errors that for example confuse class  $A_1$  with class  $A_3$  and class  $A_G$  are more significant (loss value of 2) than the ones that confuse class  $A_1$  with class  $A_2$  (loss value of 1). The main diagonal corresponds to the correct classification, so that the loss value is set to zero. This matrix can be defined in an arbitrary way, according to the situation. If it is not explicit all the errors can be assigned to have the same weight of 1.

31. The most commonly used goodness-of-fit parameters for a CM are defined in Table 4.4. When evaluating the results of a CM, the reference situation is generally taken to be the one in which all of the objects are assigned to the class that is most represented, which. This reference condition corresponds to the absence of a model, and is therefore called the *No-Model*. Goodness-of-fit values close to the ones of the *No-Model* condition give evidence of a poor result of the classification method. The *No-Model* value is unique and independent from the classification method adopted. Other statistics have been proposed, like kappa ( $k$ ) statistic (59**Error! Reference source not found.**).

32. In the particular case of a two-group CM, which evaluates the presence or absence of activity, Cooper statistics can be calculated from a 2x2 contingency table (see Table 4.5).

33. The statistics in Table 4.6 collectively express the performance of a CM, provided they measure its ability to detect known active compounds (sensitivity), non-active compounds (specificity), and all chemicals in general (concordance or accuracy). The false positive and false negative rates can be calculated from the complement of specificity and sensitivity, respectively. The positive and negative “predictivities” focus on the effects of individual chemicals, since they act as conditional probabilities. Thus, the positive “predictivity” is the probability that a chemical classified as active is really active, while the negative “predictivity” gives the probability that a classified non-active chemical is really non-active.

34. A high value of sensitivity is usually associated with a high false positive rate, so that the CM is good at identifying known active compounds, but this is at the expense of over-classifying known non-active compounds. The same relationship holds for the specificity and the rate of false negatives. Given a fixed sensitivity and specificity, the positive and negative predictivities vary according to the prevalence or proportion of active chemicals in a population, i.e.  $(a+b)/N$ . Furthermore, the accuracy is influenced by the performance of the most numerous class. Therefore, CMs should not be judged according to these statistics alone.

35. For the assessment of the predictive performance of two-group CMs, the maximal classification performance achievable should be assessed on the basis of the quality of the predictor and response data and taking also into account the purpose of the model. Thus, for stand-alone classification models, the Cooper statistics should be significantly greater than 50%, whereas for a CM that identifies active or inactive chemicals in a battery of models, a lower performance could still be useful.

36. The classification ability of a CM depends on the particular data set of chemicals used. It is therefore useful to report some measure of the variability associated with the classifications. This indicates whether the classification performance of the CM would vary significantly if it had been assessed with a different set of chemicals. To estimate the confidence intervals (CI) for the Cooper statistics, the bootstrap re-sampling technique can be used (60, 61).

37. To compare the performances of a number of classification models, the Receiver Operating Characteristic (ROC) curve can be used. ROC curves are so-named because they were first used for the detection of radio signals in the presence of noise in the 1940s (62). In the ROC graph, the X-axis is 1-specificity (false positive rate) and the Y-axis is the sensitivity (true positive rate). The best possible CM would yield a point located in the upper left corner of the ROC space, i.e. high true positive rate and low false positive rate. A CM with no discriminating power would give a straight line at an angle of 45 degrees from the horizontal, i.e. equal rates of true and false positives (63, 64). An index of the goodness of the CM is the area under the curve: a perfect CM has area of 1.0, whilst a non-discriminating test (one which falls on the diagonal) has an area of 0.5.

38. In the case of a CM based on continuous predictors, the ROC curve allows us to explore the relationship between the sensitivity and specificity resulting from different thresholds (cut points), thus allowing an optimal threshold to be determined (Figure 4.1). The threshold is an arbitrary cut-off value that determines when the prediction is considered as positive or negative. Ideally, both sensitivity and specificity would be equal to one, but changing the threshold to increase one statistic usually results in a decrease in the other. Points greater than the threshold are classified positive, whereas points less than the threshold

are classified negative. If the threshold is increased, the false positive rate decreases. However, as the false positive rate decreases, the true positive rate also decreases; this corresponds to points in the bottom left of the ROC curve. Otherwise, if the threshold is decreased, the proportion of true positives (Y axis) increases, rather dramatically initially, and then more gradually; this corresponds to points in the top right of the ROC curve.

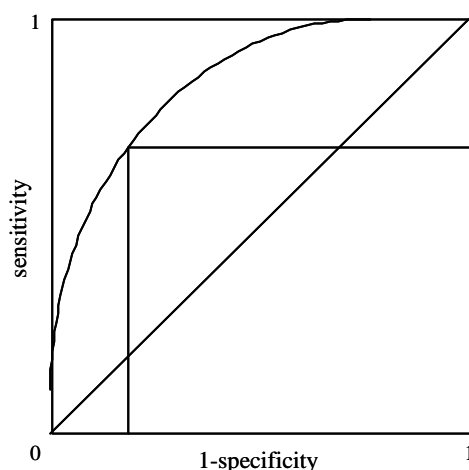


Figure 4.1 ROC curve for a model producing a continuous output. The coordinates are indicative of the performance of the models corresponding to: (0,0) high threshold, (1,1) low threshold, (0,1) perfect classification,  $y=x$ , model with no discriminatory power.

39. **Setting the importance of misclassifications.** The assessment of classification accuracy often assumes equal costs of false positives and false negative errors. However, in real applications, the minimisation of costs should be considered alongside the maximization of accuracy. The problems of unequal error costs and uneven class distributions are related, so that high-cost cases can be compensated by modifying their prevalence in the set (65).

40. The robustness of a CM can be evaluated by the total number of misclassifications, estimated with the *leave-one-out* method (66). Alternatively, the above-mentioned set of optimal loss factors (i.e. weights for different kinds of misclassifications that are minimised in the process of fitting a model) can be defined to reflect that some classification errors are more detrimental than others. The loss function represents a selected measure of the discrepancy between the observed data and data predicted by the fitted function. It can be empirically estimated and employed in a minimum risk decision rule rather than a minimum error probability rule. Also, by combining different predictions, the resulting models are more robust and accurate than single model solutions.

### ***Artificial Neural Networks (ANNs)***

41. An Artificial Neural Network (ANN) is a mathematical model that “learns” from data in a manner that emulates the learning pattern in the human brain. The calculations in a neural network model occurs as a result of the “activation” of a series of neurons, which are situated in different layers, from the input layer through one or more hidden layers to the output layer. The neural network learns by repeatedly passing through the data and adjusting its connection weights to minimise the error.

42. There are two main groups of ANN, which differ in architecture and in learning strategy: (i) unsupervised and supervised self organizing maps; and (ii) supervised back-propagation ANN (67). The terms “unsupervised” and “supervised” indicate whether only descriptors (input variables), or both descriptors and biological activities (output variables), participate in the training of ANN.

43. ANNs are especially suitable for modelling non-linear relationships and trends and have been used to tackle a variety of mathematical problems, including data exploration, pattern recognition, the modelling of continuous and categorized responses, and the modelling of multiple responses (68, 69), the classification of objects toxicological classes or modes of toxic action (71), selection of relevant descriptors, and division of the original data set into clusters (72).

### ***Measures of goodness-of-fit in neural networks***

44. Several tests for assessing the goodness-of-fit of NN models (based on the training set) are recommended. In the recall ability test (73, 74, 75, 76), the activity values are calculated for the objects of training set, to provide an indication of how well the model recognises the objects of training set. The test results are usually reported as the standard deviation and the parameters of the regression line between reference values and predicted values. Since the recall ability test is a test for goodness-of-fit only, it is recommended additional tests are also used, such as leave-one-out, leave-many-out, Y-Scrambling, and assessment with independent test set.

### ***Measures of robustness***

45. The aim of validation techniques is thus to find a model which represents the best trade-off between the model simplicity and its variability, in order to minimize the Mean Squared Error (MSE) (Table 4.7), minimising the bias as well as the unexplained variance.

46. A necessary condition for the validity of a regression model is that the multiple correlation coefficient  $R^2$  is as close as possible to one and the standard error of the estimate  $s$  small. However, this condition (*fitting ability*), which measures how well the model is able to mathematically reproduce the end point data of the training set, is an insufficient condition for model validity. In fact, models that give a high fit (smaller  $s$  and larger  $R^2$ ) tend to have a large number of predictor variables (51). These parameters are measures of the quality of the fit between predicted and experimental values, and do not express the ability of the model to make reliable predictions on new data.

47. It is well known that increasing the model complexity always increases the multiple correlation coefficient ( $R^2$ ), i.e. the explained variance in fitting, but if model complexity is not well supervised then the predictive power of the model, i.e. the explained variance in prediction ( $Q^2$ ) decreases. The differing trends of  $R^2$  and  $Q^2$  with an increasing number of predictor variables is illustrated in Figure 2.

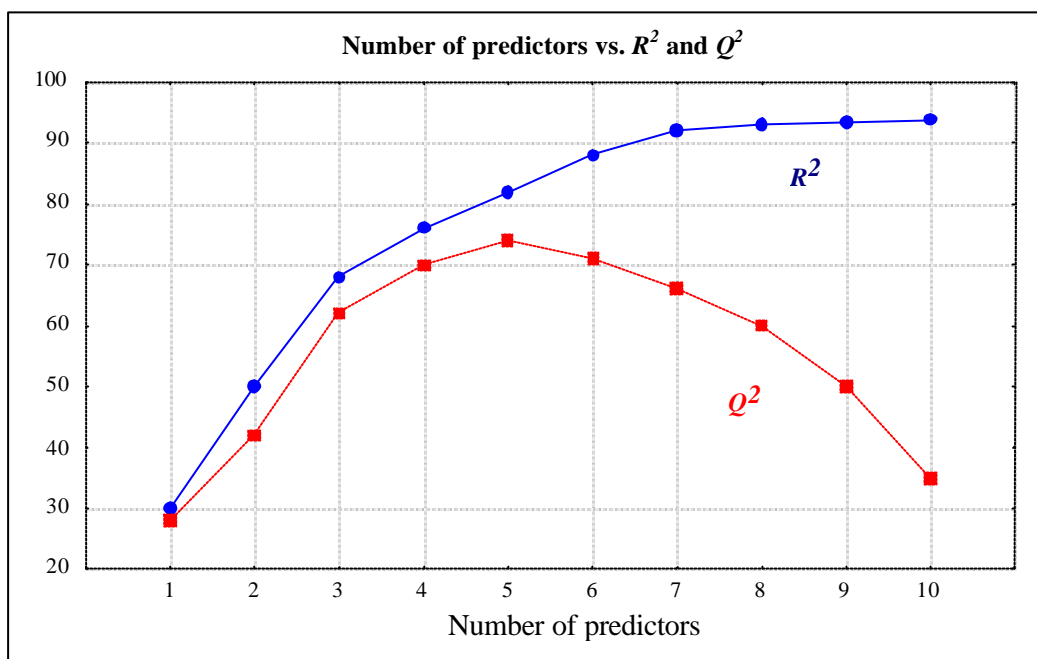


Figure 4.2. Comparison of the explained variance in fitting with the explained variance in prediction.

48. In Figure 4.2, it can be seen that increasing the number of predictors improves the explained variance in fitting ( $R^2$ ). On the other hand, the explained variance in prediction ( $Q^2$ ) only up to 5 predictors (which represents the maximum predictive power in this case) but adding further statistically insignificant predictors decreases the model performance in prediction.

49. The first condition for model validity deals with the ratio of the number of objects (i.e. chemicals) over the number of selected variables. This is called the Topliss ratio. As a rule-of-thumb, it is recommended that the Topliss ratio should have a value of at least 5.

50. The quality of multivariate regression models is usually evaluated by different fitness functions (e.g. adjusted  $R^2$ ,  $Q^2$ ) (Table 4.7) able to find the optimal model complexity and useful to compare the quality of different QSAR models.

51. For this reason, the structure of a QSAR model (number of predictors, number of PCs, number of classes) should always be inspected by validation techniques, able to detect overfitting due to variable multicollinearity, noise, sample specificity, and unjustified model complexity.

52. Model validation can be performed by internal validation techniques and external validation techniques. As illustrated in Figure 4.3, in case of internal validation a number of modified data sets are created by deleting, in each case, one or a small group of objects and each reduced data set is used to estimate the predictive capability of the final model built by using the whole data set. This means that the model predictivity is estimated by compounds (the test set) that took part in the model development, thus the information of these compounds is included in the final model. On the other hand a more demanding evaluation is the one provided by an external validation where the model predictivity is estimated by new experimentally tested compounds (external test set) that did not take part in the model development.

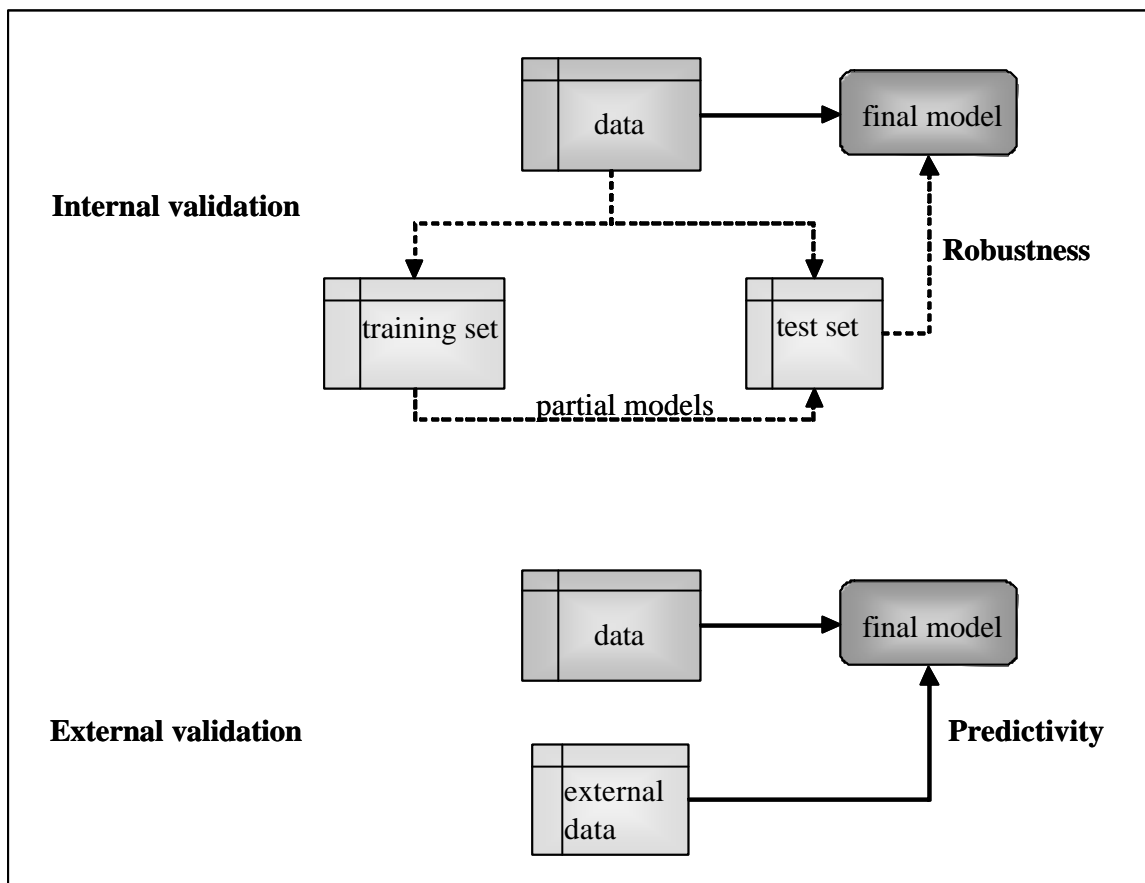


Figure 4.3. Internal and external validation.

53. A number of internal validation techniques can be used to simulate the predictive ability of a model (77, 78). The most popular validation ones are listed below:

- Cross validation (leave-one-out (LOO) and leave-many-out (LMO)).
- Bootstrapping
- Y-scrambling or response permutation testing
- Training/test set splitting

54. *Cross validation* is the most common validation technique where a number of modified data sets are created by deleting, in each case, one or a small group of compounds from the data in such a way that each object is removed away once and only once. From the original data set, a reduced data set (training set) is used to develop a partial model, while the remaining data (test set) are used to evaluate the model predictivity (79, 80). For each reduced data set, the model is calculated and responses for the deleted compounds are predicted from the model. The squared differences between the true response and the predicted response for each compound left out are added to the predictive residual sum of squares (*PRESS*). From the final predictive residual sum of squares, the  $Q^2$  (or  $R^2_{CV}$ ) and *SDEP* (*standard deviation error of prediction*) values are calculated (81) (Table 4.7).

55. The simplest cross validation procedure is the *leave-one-out* (LOO) technique, where each compound is removed, one at a time. In this case, given  $n$  compounds,  $n$  reduced models are calculated, each of these models is developed with the remaining  $n-1$  compounds and used to predict the response of the deleted compound. The model predictive power is then

calculated as the sum of squared differences between the observed and estimated response. This technique is particularly important as this deletion scheme is unique and the predictive ability of the different models can be compared accurately. However, the predictive ability obtained is often too optimistic, particularly with larger datasets compounds, because the perturbation in the dataset is small and often insignificant when only one compound is omitted.

56. To obtain more realistic estimates of the predictive ability, it is often necessary to remove more than one compound at each step. In the *leave-many-out* (LMO) cross-validation procedure, the data set is divided into a number of blocks (cancellation groups) defined by the user. At each step, all the compounds belonging to a block are left out from the derivation of the model. The cancellation groups  $G$  range from 2 to  $n$ . For example, given 120 compounds ( $n = 120$ ), for 2, 3, 5, 10 cancellation groups  $G$ , at each time  $m (= n/G)$  objects are left in the test sets, i.e. 60, 40, 24, and 12 compounds, respectively. Rules for selecting the group of compounds for the test set at each step must be adopted in order to leave out each compound only one time. The LOO method is equivalent to a LMO method with  $G = n$ , i.e. with a number of cancellation groups equal to the number of compounds. By introducing a larger perturbation in the data set, the predictive ability estimated by LMO is more realistic than the one by LOO.

57. *Bootstrap resampling* is another technique to perform internal validation (**Error! Reference source not found.**). The basic premise of bootstrap resampling is that the data set should be representative of the population from which it was drawn. Since there is only one data set, bootstrapping simulates what would happen if the samples were selected randomly. In a typical bootstrap validation,  $K$  groups of size  $n$  are generated by a repeated random selection of  $n$  compounds from the original data set. Some of these compounds can be included in the same random sample several times, while other compounds will never be selected. In this validation technique, the original size of the data set ( $n$ ) is preserved by the selection of  $n$  compounds with repetition. In this way, the training set usually consists of repeated compounds and the test set of the compounds left out (82). The model is derived by using the training set and responses are predicted by using the test set. All the squared differences between the true response and the predicted response of the compounds of the test set are expressed in the *PRESS* statistic. This procedure of building training sets and test sets is repeated thousands of times. As with the LMO technique, a high average  $Q^2$  in bootstrap validation is indicative of model robustness and what is sometimes referred to as “internal predictivity”.

58. *Y-scrambling* or response permutation testing is another widely used technique to check the robustness of a QSAR model, and to identify models based on chance correlation, i.e. models where the independent variables are randomly correlated to the response variables. The test is performed by calculating the quality of the model (usually  $R^2$  or, better,  $Q^2$ ) randomly modifying the sequence of the response vector  $y$ , i.e. by assigning to each compound a response randomly selected from the true set of responses (83). If the original model has no chance correlation, there is a significant difference in the quality of the original model and that associated with a model obtained with random responses. The procedure is repeated several hundred of times.

59. Models based on chance correlation can be detected by using the QUIK rule. Proposed in 1998 (84), the QUIK rule is a simple criterion that allows the rejection of models with high

predictor collinearity, which can lead to chance correlation (85). The QUIK rule is based on the  $K$  multivariate correlation index (Table 4.7) that measures the total correlation of a set of variables. The rule is derived from the evident assumption that the total correlation in the set given by the model predictors  $X$  plus the response  $Y$  ( $K_{XY}$ ) should always be greater than that measured only in the set of predictors ( $K_X$ ). Therefore, according to the QUIK rule only models with the  $K_{XY}$  correlation among the  $[X + Y]$  variables greater than the  $K_X$  correlation among the  $[X]$  variables can be accepted. The QUIK rule has been demonstrated to be very effective in avoiding models with multi-collinearity without prediction power.

60. An example of the application of the QUIK rule in QSAR studies is provided (84) by a series of 11 3-quinuclidinyl benzylates represented by three physicochemical descriptors: Norrington's lipophilic substituent constant  $\pi_N(x_1)$ , its squared values  $\pi_N^2(x_2)$ , and the Taft steric constant  $E_s(x_3)$ . The  $y$  response was the apparent equilibrium constant  $K_{app}$ . This data set has been extensively discussed by Stone & Jonathan (86) and by Mager (87), who concluded that the model has multicollinearity without prediction power. The regression model obtained by Ordinary Least Squares regression (OLS) was:

$$y = -8.40 + 8.35 x_1 - 1.70 x_2 + 1.43 x_3 \quad (\text{Eq 1})$$

with the following statistics:

$$R^2 = 91.8 \quad Q^2_{LOO} = 81.5 \quad Q^2_{LMO} = 67.0$$

where  $R^2$ ,  $Q^2_{LOO}$  and  $Q^2_{LMO}$  are the explained variances in fitting, by leave-one-out cross validation and by leave-many-out cross validation (two objects left out at each step), respectively. The large decrease in the predictive performance of the model was already suspect. The same conclusions were reached applying the QUIK rule. In fact, for the proposed model, the  $K$  values were:

$$K_{xy} = 47.91 < K_x = 54.87$$

According to the QUIK rule, the model would be rejected, the  $X$  correlation being greater than the  $X+Y$ - correlation.

61. Another method to check chance correlation is to add a percentage of artificial noisy variables to the set of available variables. This approach allows the detection of optimal model size, i.e. the size for which no noisy variable is present in models of this size and an example of its capability tested on a spectral matrix was extensively illustrated in Jouan-Rimbaud *et al.* (38). In fact, when simulated noisy variables start to appear in the evolving model population it means that the allowed maximum model size can no longer be increased since optimal complexity has been reached. However, this approach does not account for the likely correlation between a generated noisy variable and the  $Y$  response. In fact, there is a high probability that on generating a number of noisy predictors some will be significantly correlated with the  $y$  response. While chance correlation is considered explicitly in the  $Y$  scrambling procedure, by response randomisation, a noisy predictor could play an important role in modelling in the latter approach, contributing in the same way as a true predictor with a small, but significant, correlation with response. For this reason, noisy variables should only be used if a check on their correlation with the  $y$  response is performed first, excluding all the noisy variables with correlation greater than a fixed threshold value (85). An optimal value of this threshold can be chosen only if the experimental error of the response is known *a priori*.



62. The training/test set splitting is a validation technique based on the splitting of the data set into a training set and an test set. The model is derived from the training set and the predictive power is estimated by applying the model to the test set. The splitting is performed by randomly selecting the objects belonging to the two sets. As the results are strongly dependent on the splitting of the data, this technique is better used by repeating the splitting several hundred of times and averaging the predictive capabilities, i.e. using the repeated test set technique (88).

### ***Measures of predictivity***

63. One of the most important characteristics of a (Q)SAR model is its predictive power, i.e. the ability of a model to predict accurately the (biological) activity of compounds that were not used for model development. While the internal validation techniques described above can be used to establish model robustness, they do not directly assess model predictivity.

64. In principle, external validation is the only way to “determine” the true predictive power of a QSAR model. This type of assessment requires the use of an external test set, i.e. compounds not used for the model development. It is generally considered the most rigorous validation procedure, because the compounds in the external test set do not affect the model development. In fact, the test set is often constituted of new experimentally tested compounds used to check the predictive power of the model.

65. External validation should be regarded as a supplementary procedure to internal validation, rather than as a (superior) alternative. This is because a model that is externally predictive should also be robust, although a robust model is not necessarily predictive (of independent data). Indeed, a high value of the leave-one-out cross-validated correlation coefficient,  $Q^2$ , can be regarded as a necessary, but insufficient, condition for a model to have a high predictive power (89).

66. The predictivity of a regression model is estimated by comparing the predicted and observed values of a *sufficiently large and representative* external test set of compounds that were not used in the model development. By using the selected model, the values of the response for the test objects are calculated and the quality of these predictions is defined in terms of external explained variance  $Q^2_{ext}$  (Table 4.7). Unlike the cross-validated correlation coefficient,  $Q^2$ , in the external explained variance  $Q^2_{ext}$  the sum of the predictive residual sum of squares on the numerator runs over the external test chemicals and the reference total sum of squares on the denominator is calculated comparing the predicted response of the external test chemicals with the average response of the training set.

67. Analogously, the predictivity of a classification model is estimated by comparing the predicted and observed classes of a *sufficiently large and representative* test set of compounds that were not used in the model development. The parameters described in Table 4.4, but derived by using the external test set, are used to quantify the CM predictivity.

68. In practice, for reasons of cost, time and animal welfare, it is often difficult or impossible to obtain new experimentally tested compounds to check model predictivity, and for this reason a common practice is to split the available dataset into training set, used to develop the (Q)SAR model and an external test set, containing compounds not present in the

training set and used to assess the predictive capability of a (Q)SAR model. This technique can be used reliably only if the splitting is performed by partitioning the compounds according to a pre-defined and suitable criterion, such as a criterion based on experimental design or cluster analysis.

69. When performing statistically designed external validation, the goal is to ensure that: a) the training and test sets separately span the whole descriptor space occupied by the entire data set; and b) the structural domains in the two sets are not too dissimilar. It is important that the training set contains compounds that are informative and good representatives of many other similar compounds. Thus, the following criteria were recently proposed for training and test selection (90): a) representative points of the test set must be close to points in of the training set; b) representative points of the training set must be close to points in the test set; and c) the training set must be diverse. These criteria were proposed to ensure that the similarity principle can be adopted when predicting the test set.

70. To accomplish a well-planned selection, some approach to statistical experimental design is needed (91). An ideal splitting leads to a test set such that each of its members is close to at least one member of the training set (92). Developing rational approaches for the selection of training and test sets is an active area of research. These approaches range from the straightforward random selection (93) through activity sampling and various systematic clustering techniques (94, 95), to the methods of self-organising maps (96), Kennard Stone (97), formal statistical experimental design (factorial and D-Optimal) (98), and recently proposed modified sphere exclusion algorithm (99). These methods help achieve desirable statistical characteristics of the training and test sets.

71. A frequently used approach is *activity sampling* (100), according to which the choice of training and test sets is made by binning the range of experimental values and randomly selecting an even distribution of compounds from each bin. This guarantees that members of the test set span the entire range of the experimental measurements and are numerically representative of the data set. However, because the binning is based on the response, it does not guarantee that the training set represents the entire descriptor space of the original dataset and that each compound of the test set is close to at least one of the training set.

72. In several applications, the training/test splitting is performed by using clustering techniques. *K*-means algorithm is often used, and from each cluster one compound for the training set is randomly selected. Given that all compounds are represented in a multidimensional descriptor space, the clustering algorithm can be performed on the descriptor values (X values), on response values (Y values), or on the descriptor/response values (X/Y values). Clustering on X/Y values allows clustering the compounds according to all of the given information (101). An alternative clustering approach to select representative subset of compounds is the one based on the maximum dissimilarity method (94, 95). The method starts with the random selection of a seed compound, then every new compound is successively selected such that it is maximally dissimilar from all the other compounds of the dataset. The process ends either when a maximum number of compounds has been selected or when no other compound can be selected without being too similar to one already selected. Since this method is based on a random starting point, the variance of the results is normally checked by comparing various selections. Hierarchical clustering provides a more specific control by assigning every single compound to a cluster of compounds. It does not require any prior assumption about the number of clusters, and after the clustering process the compound closest to the centre of a cluster is selected as representative compound.

73. Another way to perform a statistically planned training/test selection is by using the Kohonen's Self-Organising Maps (102). The main goal of the neural network is to map compounds from  $n$ -dimensional into two-dimensional space. Representative compounds falling in the same areas of the map are randomly selected for the training and test sets. This approach preserves the closeness between compounds: compounds which are similar in the original multidimensional space are close to each other on the map.

74. Similarly to the maximum dissimilarity method, the Kennard Stone algorithm can be used to perform data splitting (103). It is sequential and consists in maximizing the Euclidean distances between the newly selected compounds and the ones already selected. An additional compound is selected by calculating for each compound, which is not selected, the distance to each selected compound and by maximizing the distance to the closest compound already selected. Both the maximum dissimilarity and the Kennard Stone methods guarantee that the training set compounds are distributed more or less evenly within the whole area of the representative points, and the condition of closeness of the test set to the training set is satisfied.

75. Another data splitting strategy makes use of fractional factorial design (FFD) and D-Optimal design (factorial and D-Optimal) (98). A common practice is to process the original data using principal component analysis (PCA) and subsequently to use the principal components (PCs) as design variables in a design selecting a small number of informative and representative training data. These principal components are suitable for experimental design purposes since they are orthogonal and limited in number, reducing the extent of collinearity in the training set. In fractional factorial design, all the principal components are explored at two, three or five levels. The training set includes one representative for each combination of components. The drawback of this approach is that it does not guarantee the closeness of the test set to the training set in the descriptor space. D-Optimal design is often performed whenever the classical symmetrical design cannot be applied, because the experimental region is not regular in shape or the number of compounds is selected by a classical design is too large. The basic principle of this method is to select compounds to maximize the determinant of the information (variance-covariance) matrix  $|X'X|$  of independent variables. The determinant of this matrix is maximal when the selected compounds span the space of the whole data, i.e. when the most influential compounds (maximal spread) are selected.

76. Sphere Exclusion is a dissimilarity-based compound selection method first described by Hudson *et al.* (104) and then later adapted by various groups (99, 105, 106). The algorithm consists in selecting molecules, whose similarities with each of the other selected molecules are not higher than the defined threshold (106). Therefore, each selected molecule creates a (hyper) sphere around itself, so that any candidate molecules inside the sphere are excluded from the selection. The radius of the sphere is an adjustable parameter, determining the number of compounds selected and the diversity between them. The original method starts with the "most descriptive compound" and in each cycle identifies the compound most similar to the centroid of the already selected compounds. This was considered to be very computer intensive, so variations from the original algorithm have been implemented to reduce the computer time required by selecting the next compound quicker.

## Concluding remarks

77. Ideally, QSAR modelling should lead to statistically robust models capable of making reliable predictions for new compounds. In this Guidance Document, reference is made to the reliability, rather than the correctness, of model predictions. This is because from a philosophical viewpoint, it is questionable whether a prediction can ever be correct, or whether a model can ever truly represent reality. As famously quoted by the chemist and statistician, George Box (91), “all models are wrong, but some are useful”.

78. In order for a statistical model to be useful for predictive purposes, it should be built on a sufficiently large and representative amount of information regarding the modelled activity and should contain only relevant variables. As discussed in this chapter, a variety of statistical methods are available for assessing the goodness-of-fit, robustness and predictive ability of QSAR models, and a variety of statistics are routinely used to express these aspects of model performance. Modern statistical software packages provide convenient and automated means of applying these methods and generating a plethora of statistics. The users of (Q)SAR models, such as regulators, need a sufficient understanding of these statistics and the underlying methods in order to interpret the statistics according to their own purposes.

79. The model user should be aware that the performance of a model, while being expressed in quantitative terms and on the basis of well-established techniques, is dependent on the choices by the (Q)SAR modeller. Different types of statistics are generated by different methods, and different values of the same statistics can be generated by altering the compositions of the training and test sets, or by altering the resampling routine in a cross-validation procedure. This is why transparency in the statistical validation process is needed to form the basis of sound decision-making.

80. Internal validation refers to the assessment of goodness-of-fit and robustness. The goodness-of-fit of a model to its training set can be regarded as the absolute minimum of information needed to assess model performance. It expresses the extent to which the model descriptors “account for” the variation in the training set, and most importantly whether the model is statistically significant. If the model is not statistically significant, or if it is significant but of poor fit, it cannot be expected to be useful for predictive purposes.

81. The robustness of the model provides an indication of how sensitive the model parameters (and therefore predictions) are to changes in the training set. If the model is not robust to small perturbations in the training set, it is unlikely to be useful for predictive purposes. In practice, robustness can be a difficult concept to apply, because there are numerous ways of resampling the data, which affect the statistics generated.

82. The distinction between internal and external validation has important practical implications. Models that are too complex (i.e overfitted) are unlikely to predict independent data as reliably as their internal validation statistics may imply. This problem is increasingly relevant as modern QSAR methods become more powerful and capable of handling large amounts of correlated information and a large number of noisy variables.

83. Predictivity is perhaps the most difficult concept to apply. From a philosophical standpoint, it can be argued that it is impossible to determine an absolute measure of

predictivity, since it is highly dependent on the choice of statistical method and test set. Nevertheless, external validation, when performed judiciously, is generally regarded as the most rigorous assessment of predictivity, since predictions are made for chemicals not used in the model development.

84. External validation should be seen as a useful supplement to internal validation, rather than as a substitute. External validation can be difficult to apply in a meaningful way when data of sufficient quality are scarce. The model user should therefore be aware that the statistics derived by external validation could be less meaningful than those provided by internal validation, if the external test set is not carefully designed.

85. It is not the aim of this document to define acceptability criteria for the regulatory use of QSAR models, since the use of data in decision-making is highly context-dependent. However, it is possible to identify features of models that are likely to contribute to a high or low performance.

86. A model with high statistical performance is likely to have one or more of the following characteristics:

- a) the highest possible prediction power is achieved with the minimum number of variables;
- b) there is a low correlation between the predictor variables.

87. A model with low statistical performance is likely to have one or more of the following characteristics:

- a) it is lacking one or more relevant variables, i.e. has insufficient fitting capability;
- b) there is a marked difference between goodness-of-fit and prediction power;
- c) one or more (noisy) variables are correlated with the response by chance;
- d) there is a high correlation between the predictor variables (multi-collinearity) resulting in redundancy in descriptor information.

Table 4.1. Basic equations and parameters of goodness of fit in MLR

N.	Definition	Equation and terms
1	MLR equation	$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots b_px_p$ <ul style="list-style-type: none"> <li>• <math>\hat{y}</math> = calculated dependent variable</li> <li>• <math>x_j</math> = predictor variable</li> <li>• <math>b_j</math> = regression coefficient</li> </ul>
2	Coefficient of multiple determination (Multiple correlation coefficient )	$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{(SS_T - SS_{Res})}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$ <ul style="list-style-type: none"> <li>• <math>SS_T = \sum_i (y_i - \bar{y})^2</math> = total sum of squares</li> <li>• <math>SS_{Res} = \sum_i (y_i - \hat{y})^2</math> = residual sum of squares</li> <li>• <math>SS_{Reg} = \sum_i (\hat{y} - \bar{y})^2</math> = sum of squares due to the regression</li> <li>• <math>y_i</math> = observed dependent variable</li> <li>• <math>\bar{y}</math> = mean value of the dependent variable</li> <li>• <math>\hat{y}</math> = calculated dependent variable</li> </ul>
3	Adjusted $R^2$	$R_{adj}^2 = 1 - \frac{SS_{Res}/(n-p-1)}{SS_T/(n-1)}$ $= 1 - (1 - R^2) \cdot \left( \frac{n-1}{n-p-1} \right)$ <ul style="list-style-type: none"> <li>• <math>n</math> = number of observations</li> <li>• <math>p</math> = number of predictor variables</li> </ul>
4	Standard error of estimate	$s = \sqrt{\frac{\sum_i (y_i - \hat{y})^2}{(n-p-1)}}$
5	$F$ -value	$F = \frac{EMS}{RMS} = \frac{SS_{Reg}/p}{SS_{Res}/(n-p-1)}$ <ul style="list-style-type: none"> <li>• <math>RMS = SS_{Res}/(n-p-1)</math> = residual mean square</li> <li>• <math>EMS = SS_{Reg}/p</math> = explained mean square</li> </ul>
6	$t$ -test	$t = \frac{b_j}{s_{b_j}}$ <ul style="list-style-type: none"> <li>• <math>s_{b_j} = \frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}} = \frac{MS_R}{\sqrt{\sum_i (x_i - \bar{x})^2}}</math> = standard deviation of the estimated regression coefficient <math>b_j</math></li> <li>• <math>\bar{x}</math> = mean value of the predictor variable</li> </ul>

Table 4.2. Confusion or contingency matrix  $\{c_{GG}\}$  for a general case with  $G$  classes

		Assigned class				Marginal totals
		$A1'$	$A2'$	$A3'$	$A_g'$	
True class	$A1$	$c_{11'}$	$c_{12'}$	$c_{13'}$	$c_{1g'}$	$n_1$
	$A2$	$c_{21'}$	$c_{22'}$	$c_{23'}$	$c_{2g'}$	$n_2$
	$A3$	$c_{31'}$	$c_{32'}$	$c_{33'}$	$c_{3g'}$	$n_3$
	$A_g$	$c_{G1'}$	$c_{G2'}$	$c_{k3'}$	$C_{gg'}$	$n_g$
Marginal totals		$n_{1'}$	$n_{2'}$	$n_{3'}$	$n_{g'}$	

Table 4.3. Example of loss matrix  $\{l_{GG}\}$  where the loss function has been arbitrarily defined in an integer scale

		Assigned class			
		$A1'$	$A2'$	$A3'$	$A_g'$
True class	$A1$	0	1	2	2
	$A2$	1	0	1	1
	$A3$	2	1	0	2
	$A_g$	2	1	2	0

Table 4.4. Definitions of the goodness-of-fit parameters

Statistic	Formula	Definition
Concordance or Accuracy (Non-error Rate)	$\frac{\sum_g c_{gg'}}{n} \times 100$	total fraction of objects correctly classified.  $c_{gg'}$ = number of objects correctly classified to each class $n$ = total number of objects
Error Rate	$\frac{n - \sum_g c_{gg'}}{n}$ 1-concordance	total fraction of objects misclassified  $c_{gg'}$ = number of objects correctly classified to each class $n$ = total number of objects Error provided in absence of model
<b><i>NO-Model Error Rate, NOMER%</i></b>	$NOMER\% = \frac{n - n_M}{n} \times 100$	$n_M$ = number of objects of the most represented class $n$ = total number of objects
Prior probability of a class	$P_g = \frac{1}{G}$	probability that an object belongs to a class supposing that every class has the same probability  $G$ = number of classes
Prior proportional probability of a class	$P_g = \frac{n_g}{n}$	probability that an object belongs to a class taking into account the number of objects of the class  $n_g$ = total number of objects belonging to class $g$ $n$ = total number of objects
<b>Sensitivity of a class</b>	$\frac{c_{gg'}}{n_g} \times 100$	percentage of objects correctly assigned to the class out of the total number of objects <i>belonging</i> to that class  $c_{gg'}$ = number of objects correctly classified to each class $n_g$ = total number of objects belonging to class $g$
<b>Specificity of a class</b>	$\frac{c_{gg'}}{n_{g'}} \times 100$	percentage of objects correctly assigned to the class out of the total number of objects <i>assigned</i> to that class  $c_{gg'}$ = number of objects correctly classified to each class $n_{g'}$ = total number of objects assigned to class $g$



<b>Misclassification risk</b>	$\sum_g \frac{(\sum_{g'} l_{gg'} c_{gg'}) P_g}{n_g} \times 100$	<p>risk of incorrect classification (takes into account the number of missclassifications, and their importance)</p> <p><math>c_{gg'}</math> = number of objects correctly classified to each class</p> <p><math>n_g</math> = total number of objects belonging to class <math>g</math></p> <p><math>P_g</math> = <b>prior probability class</b></p>
-------------------------------	---	--

Footnote:

$g=1, \dots, G$  ( $G$  = number of classes)

Table 4.5.  $2 \times 2$  contingency table

		Assigned class		Marginal totals
		Toxic	Non-toxic	
<b>Observed (<i>in vivo</i>) class</b>	<b>Active</b>	$a$	$b$	$a+b$
	<b>Non-active</b>	$c$	$d$	$c+d$
	<b>Marginal totals</b>	$a+c$	$b+d$	$a+b+c+d$

Table 4.6. Definitions of the Cooper statistics

Statistic	Formula	Definition
<b>Sensitivity</b> (True Positive rate)	$a/(a+b)$	fraction of active chemicals correctly assigned
<b>Specificity</b> (True Negative rate)	$d/(c+d)$	fraction of non-active chemicals correctly assigned
<b>Concordance or Accuracy</b>	$\frac{(a+d)}{(a+b+c+d)}$	fraction of chemicals correctly assigned
<b>Positive Predictivity</b>	$a/(a+c)$	fraction of chemicals correctly assigned as active out of the active assigned chemicals
<b>Negative Predictivity</b>	$d/(b+d)$	fraction of chemicals correctly assigned as non-active out of the non-active assigned chemicals
<b>False Positive (over-classification) rate</b>	$\frac{c}{(c+d)}$ <b>1-specificity</b>	fraction of non-active chemicals that are falsely assigned to be active
<b>False Negative (under-classification) rate</b>	$\frac{b}{(a+b)}$ <b>1-sensitivity</b>	fraction of active chemicals that are falsely assigned to be non-active

Table 4.7. Definitions of the robustness and predictive parameters

Statistic	Definition	Formula
$MSE$	Mean Squared Error	$MSE(\mathbf{b}) = E[(\mathbf{b} - \mathbf{b})^2] = E[(\mathbf{b} - E(\mathbf{b})) - (\mathbf{b} - E(\mathbf{b}))]^2 = V(\mathbf{b}) + B^2(\mathbf{b})$ <p> <math>\mathbf{b}</math> = estimator of the true value <math>\beta</math>  <math>\mathbf{b}</math> = true value of a quantity  <math>E(\mathbf{b})</math> = expected value of <math>\mathbf{b}</math>  <math>V(\mathbf{b})</math> = Variance of the estimator <math>\mathbf{b}</math>  <math>B^2(\mathbf{b})</math> = bias of the estimator <math>\mathbf{b}</math>  <math>PRESS = \sum_i (y_i - \hat{y}_{i/i})^2</math> </p>
$PRESS$	Predictive Residual Sum of Squares	<p> <math>y_i</math> = observed response for the <math>i</math>-th object  <math>\hat{y}_{i/i}</math> = response of the <math>i</math>-th object estimated by using a model obtained without using the <math>i</math>-th object </p> $Q^2 = 1 - \frac{PRESS}{SS_T} = 1 - \frac{\sum_i (y_i - \hat{y}_{i/i})^2}{\sum_i (y_i - \bar{y})^2}$
$Q^2$	Explained variance in prediction	<p> <math>SS_T</math> = total sum of squares  <math>y_i</math> = observed response for the <math>i</math>-th object  <math>\hat{y}_{i/i}</math> = response of the <math>i</math>-th object estimated by using a model obtained without using the <math>i</math>-th object  <math>\bar{y}</math> = average response value of the training set </p> $SDEP = \sqrt{\frac{\sum_i (y_i - \hat{y}_{i/i})^2}{n}}$
$SDEP$	Standard Deviation Error of Prediction	<p> <math>y_i</math> = observed response for the <math>i</math>-th object  <math>\hat{y}_{i/i}</math> = response of the <math>i</math>-th object estimated by using a model obtained without using the <math>i</math>-th object  <math>n</math> = the number of training objects </p>
$K$	Multivariate correlation index	$K\% = \frac{\sum_m \left  \frac{I_m}{\sum_m I_m} - \frac{1}{p} \right }{\frac{2(p-1)}{p}} \times 100$ <p> <math>?m</math> = eigenvalues obtained from the correlation matrix of the data set <math>\mathbf{X}(n, p)</math>,  <math>n</math> = number of objects  <math>p</math> = number of variables. </p>
$Q_{ext}^2$	External explained variance	$Q_{ext}^2 = 1 - \frac{\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \bar{y})^2}$

---

$y_i$  = observed response for the  $i$ -th object

$\hat{y}_i$  = predicted response for the  $i$ -th object

$\bar{y}$  = average response value of the training set

---



## **CHAPTER 5**

### **MECHANISTIC RELEVANCE**

## CHAPTER 5: MECHANISTIC RELEVANCE

### Summary of chapter 5

This chapter provides guidance on the application of the principle, “a (Q)SAR should be associated with a mechanistic interpretation, if possible”. The chapter begins with a historical perspective citing several early examples of congeneric (Q)SAR models where the notion of mechanistic interpretation first began. It then goes on to describe examples of more recent (Q)SARs where mechanistic interpretations have been provided. The difference between what is meant by a mechanistic basis and a mechanistic interpretation is clarified through the use of these examples. The chapter also makes raises several discussion points and proposes potential areas for further research.

### Introduction

1. A mechanistic understanding of a (Q)SAR can add to the confidence in the model already established on the basis of its transparency and predictive ability. For this reason, a mechanistic understanding can contribute to the acceptance and use of a (Q)SAR. The intent of this principle is therefore to ensure that an assessment of the mechanistic association between the descriptors used in the model and the endpoint being predicted has at least been considered, and if there is a plausible mechanistic association, that this has been documented. However, depending on the intended application of the (Q)SAR, this kind of information could be regarded as unimportant, desirable or essential.

2. A *molecular descriptor* is a structural or physicochemical property of a molecule, or part of a molecule, which characterises a specific aspect of a molecule and is used as an independent variable in a QSAR. According to their physicochemical interpretation, descriptors are often classified into three general types (electronic, steric and hydrophobic). Table 5.1 provides a list of commonly used descriptors in (Q)SAR studies.

3. It should be noted that a model based on transparent descriptors and on a transparent algorithm can always be given an interpretation, but the challenge is to assess the plausibility or likelihood of that interpretation. Such an assessment can only be made on a model-by-model basis. Such an the assessment will depend partly on the state of knowledge in the field (i.e. knowledge of the features captured by the model descriptors and the (biological) relevance of the association between the descriptors and the endpoint. The assessment will also be partly subjective, depending on the training and experience of the individual. For this reason, the principle will be difficult to apply in a consistent way between different experts.

4. From a philosophical standpoint, it is interesting to distinguish between the terms “mechanistic basis” and “mechanistic interpretation”. The term *mechanistic basis* refers to the development of a mechanistic hypothesis before modelling (*a priori*). The mechanistic hypothesis is proposed to determine which factors/properties are likely to be significant and thus which descriptors will be best to model that property. In such cases, it is important to use descriptors with a clear or widely accepted physicochemical interpretation. In contrast, *mechanistic interpretation* refers to the assignment of physical/chemical/biological meaning

to the descriptors after modelling (*a posteriori*). Thus, the descriptors used in the final model are interpreted as to their relevance to the toxicity under consideration.

5. In practice, it can be difficult to distinguish between *a priori* and *a posteriori* interpretation of a model, because the modelling process often involves multiple iterations, starting from an exploratory data-mining step in which large numbers of descriptors are screened for predictive value, after which combinations of “promising” predictors are investigated, with the possible exclusion of descriptors that are considered to contain the same or similar information as other descriptors.

## Historical background

6. The basic underlying principle of (Q)SARs is that the properties and biological interactions of a chemical with a defined system are inherent in its molecular structure. Attempts to develop (Q)SARs consist of looking for links between the structure and biological activity. These links may be mechanistically based or may be purely empirical. Ideally the activities and properties are connected by some known mathematical function, F:

$$\text{Biological activity} = F(\text{Physicochemical Properties}) \quad (\text{Eq 1})$$

7. Around 1935, Hammett (108) made an enormous contribution to our ability to elucidate organic and eventually biochemical and biological reaction mechanisms. His reasoning was that similar changes in structure should produced similar changes in reactivity. Although original at the time, this reasoning is still used by organic chemists today. He postulated that the effect of substituents on the structure of benzoic acids could be used as a model system to estimate the electronic effects of substituents on similar reaction systems. The more electron attracting the substituent, the more rapid the reaction. Hammett defined a parameter, *s*. Positive values of *s* represented electron withdrawal by the substituent from the aromatic ring and negative values indicated electron release.

8. Although the Hammett equation has been modified and extended, *s* constants still remain the most general means for estimating the electronic effects of substituents on reaction centres. The power of these simple *s* values is that they often take into account solution effects on substituents such as hydrogen bonding, dipole interactions and so on that are still difficult to calculate.

9. Hammett’s reasoning was subsequently extended to the development of steric and hydrophobic parameters. These extensions have enabled all kinds of structure-activity relationships of chemical reactions to be tackled.

10. However the field of QSAR really started to flourish with the pioneering work of Corwin Hansch in the early 1960s. He proposed a mathematical model which correlated biological activity with chemical structure. He correlated the plant growth regulatory activity of phenoxyacetic acids to Hammett constants and partition coefficients (109). In 1964, Hansch *et al.* showed that the biological activity could be correlated linearly by free-energy related parameters (110). This approach became known as a Linear Free Energy Relationship (LFER) and expressed in the following equation:

$$\log 1/C = a_p + b_s + cE_s + \dots + \text{constant} \quad (\text{Eq 2})$$

In Equation 2, C is the molar concentration of the compound to produce a defined biological response, p is the hydrophobic contribution of the substituent and represented by  $\log P_X/P_H$ , s is the Hammett electronic descriptor of the substituents (111), represented by  $\log K_X/K_H$ ,  $E_S$  is Taft's steric parameter (112) and a, b and c are the appropriate coefficients. In these expressions  $P_X$  and  $P_H$  are the octanol/water partition coefficients of the substituted and unsubstituted compounds, respectively, and  $K_X$  and  $K_H$  are the ionization constants of the meta- or para-substituted and unsubstituted benzoic acids at 25 °C, respectively.

11. Hansch recognised the importance of partition effects on any attempt to describe the properties of a biological system. The reasoning behind this lay in the recognition that in order to exert an effect on a system the compound first had to reach that site of action. Since biological systems are composed of a variety of aqueous phases separated by membranes, measurement of partition coefficients in a suitable system of immiscible solvents was thought to provide a simple chemical model of these partition steps in the biological system. Hansch chose an octanol and water system. Octanol was chosen for a variety of reasons perhaps the most important of these that it consists of a long hydrocarbon chain with a relatively polar hydroxyl head group and therefore mimics some of the lipid constituents of biological membranes. The generalized form of what has now become known as the Hansch approach is shown as

$$\log 1/C = a p + b p^2 + c s + d E_s + \text{constant} \quad (\text{Eq 3})$$

In Equation 3, C is the dose required to produce a standard effect,  $p_i$ , sigma and  $E_s$  are hydrophobic, electronic and steric parameters and a,b,c,d are statistical parameters fitted by linear regression. Const is a constant. The squared term for  $p_i$  is included in an attempt to account for non-linear relationships in hydrophobicity.

12. The work of Hansch provided perhaps the first example of how a (Q)SAR could give information concerning mechanism. He and his workers (113) demonstrated the following relationship for a set of esters binding to the enzyme papain.

$$\text{Log } 1/K_m = 1.03p_3' + 0.57s + 0.61MR_4 + 3.8 \quad (\text{Eq 4})$$

N = 25, r = 0.907, s = 0.208

In Equation 4,  $K_m$  the Michaelis-Menten constant, is the substrate concentration at which the rate of the reaction is half maximal. The subscripts to the physicochemical parameters indicate substituent positions. The statistics quoted are the number of compounds in the dataset, the correlation coefficient, a measure of the goodness-of-fit and the standard error of the fit.

13. Physicochemical meaning was assigned to the physicochemical parameters in Equation 4 as follows. The positive sigma term implied that electron withdrawing substituents favoured formation of the enzyme substrate complex. This made biological sense since the MOA of papain does involve the electron rich thiol group of a cysteine residue. The positive molar refractivity term suggested that bulkier substituents in the 4 position favoured binding. The two parameters  $p_4$  and  $MR_4$  are orthogonal to each other in the dataset implying that a bulk effect rather than a hydrophobic effect was important at position 4. The prime sign associated with the p parameter for position 3 indicated that in cases where there were two



meta substituents, the  $p$  value of more hydrophobic substituent was used, the other  $p$  3 value being ignored. The rationale for this was that binding of one meta substituent to the enzyme placed the other into an aqueous region and therefore outside the enzyme binding site (114).

14. Hansch's early work clearly demonstrates the concept of mechanistic interpretation through the assignment of physicochemical meaning to the descriptors. This approach has been widely followed by other workers. Nowadays, *Hansch analysis* refers to the investigation of the quantitative relationship between the biological activity of a series of compounds and their physicochemical substituent or global parameters representing hydrophobic, electronic, steric and other effects using multiple regression techniques.

15. Whilst the Hansch approach is mechanistically simple, it is somewhat limited in its breadth of application. Typically, Hansch-type QSARs are limited to congeneric series of chemicals i.e. groups of chemicals with a common parent structure (e.g. aliphatic alcohols). In practice, it is desirable to develop a (Q)SAR that is applicable to a wider range of chemicals than a single series of chemicals with the same backbone. Thus, optimising maximum diversity in chemical structure with a sound mechanistic basis is an important challenge (Figure 5.1).

16. The development of (Q)SARs has therefore evolved in an attempt to accommodate models which optimise the structural, data and mechanistic diversity. The two main approaches in the (Q)SAR development now are mechanistically based (Q)SARs and empirical (or correlative) (Q)SARs. These may be considered in the following way.

17. Mechanism-based (Q)SARs are those (Q)SARs where a hypothesis is made as to the physicochemical properties or descriptors that are likely to be relevant. Statistical methods are then applied to seek out correlations existing between these descriptors and the endpoint of interest.

18. In the case of purely empirical approaches, no assumptions are made as to the likely (biological) mechanism. A large number of physicochemical parameters or structural parameters are calculated and statistical approaches are applied to identify those features that correlate most closely with the biological activity.

19. In both cases a physicochemical meaning can be assigned to interpret the descriptors after modelling, the only difference lies in whether descriptors are pre-determined with a mechanism in mind. In practice, (Q)SARs are often developed in an iterative manner, with some descriptors being included *a priori* on the basis of mechanistic "expectation", and others being included *a posteriori* on the basis of mechanistic "discovery".

## **Recommendations for practitioners**

20. In this section, a few examples illustrate specific (Q)SARs where the concept of mechanistic interpretation or mechanistic basis has been applied. A range of (Q)SARs have been explored for both human health and environmental effects for endpoints, including mutagenicity, eye irritation, skin sensitisation, as well as aquatic toxicity.

## Human Health Endpoints

21. Benigni *et al.* (115) aimed to study some molecular determinants to discriminate between mutagenic and inactive compounds for aromatic and heteroaromatic amines and nitroarenes. Using a selection of data from the literature (both Ames and SOS repair), he investigated the feasibility of developing (Q)SARs. He found a dramatic difference between those (Q)SARs derived for estimating potency and those derived for predicting the absence or presence of activity. Hydrophobicity was found to play a major role in determining the potency of the active compounds whereas mainly electronic factors differentiated the actives from the inactives. The electronic factors were those expected on the basis of hypothesised metabolic pathways of the chemicals. Electronic factors together with size/shape appeared to determine the minimum requirement for the chemicals to be metabolised whereas hydrophobicity determined the extent of activity.

22. Debnath *et al.* (116) modelled mutagenic potency in the TA98 strain of *Salmonella typhimurium* (+ S9 activation system) and derived the following equation for a set of aminoarenes:

$$\log \text{TA98} = 1.08 \log P + 1.28 \text{HOMO} - 0.73 \text{LUMO} + 1.46 \text{IL} + 7.20 \quad (\text{Eq 5})$$

$n = 88, r = 0.898 (r^2 = 0.806), s = 0.860, F_{1,83} = 12.6$

The mutagenic potency ( $\log \text{TA98}$ ) was expressed as  $\log$  (revertants/nmol). *IL* in the equation was an indicator variable that assumed a value of 1 for compounds with three or more fused rings. Overall, the principal factor affecting the relative mutagenicity of the aminoarenes was their hydrophobicity ( $\log P$ ). Mutagenicity increased with increasing HOMO values; this positive correlation seemed reasonable since compounds with higher HOMO values are easier to oxidize and should be readily bioactivated. For the negative correlation with LUMO, no simple explanation could be offered.

23. Barratt (117) proposed a mechanism-based model for predicting the eye irritation potential of neutral organic chemicals, as measured in the rabbit draize eye test. A substance which is classified as irritating to eyes according to EC criteria is one which causes a defined degree of trauma in the Draize rabbit eye test following the instillation of 0.1ml (or equivalent weight) as defined in the EC Annex V method (118) and the OECD Test Guideline 405 (119). Neutral organics were described as uncharged, carbon-based chemicals which did not possess the potential to react covalently with or to ionize under the conditions prevalent in biological systems. Common chemical classes covered by this definition were hydrocarbons, alcohols, ethers, esters, ketones, amides, unreactive halogenated compounds, unreactive aromatic compounds and aprotic polar chemicals. Data on 38 neutral organics taken from the reference databank of eye irritation data published by ECETOC (European Centre for Ecotoxicology and Toxicology) (120) together with 8 chemicals drawn from work by Jacob & Martens (121) was analysed using principal components analysis (PCA). The mechanistic hypothesis underlying this (Q)SAR was summarized as follows. Neutral organic chemicals were irritant as a result of the perturbation of ion transport across cell membranes. These perturbations arise from changes in the electrical properties of the membrane and are related to dipole moments of the perturbing chemicals. In order to affect these electrical properties, a chemical must be able to partition into the membrane and hence possess the appropriate hydrophobic/hydrophilic properties. An appropriately small cross sectional area allowing it to fit easily between lipid components of the membrane was also a requirement.  $\log P$  was used as a measure of hydrophobicity. The minor principal inertial axes  $R_y$  and  $R_z$  were used to

represent the cross-sectional area and the dipole moment was used to model the reactivity. Plots of the first two principal components of these parameters showed that PCA was able to discriminate well between the irritant and non-irritant chemicals in the dataset.

24. Abraham and his workers followed a similar mechanistic based approach. In this example a collection of data on the Draize rabbit eye test was analysed (122) using the set of Abraham (123) descriptors. These descriptors included  $R_2$ , excess molar refraction,  $\pi_2^H$  polarisability/dipolarity,  $\alpha_2^H$  and  $\beta_2^H$  effective hydrogen bond acidity and basicity and  $\text{Log } L^{16}$  a descriptor where  $L^{16}$  is the vapour-hexadecane solubility at 25°C. A possible model process would be that of transfer of a pure organic liquid to a dilute solution in an organic solvent phase. The equilibrium constant governing such a model process is known as the activity coefficient,  $\gamma^\circ$ , which may be defined for a sparingly soluble liquid as the reciprocal of the solubility of the liquid in the organic solvent phase. Abraham defined the solubility of a vapour into a solvent phase as  $L$ , where  $L = (1/g^\circ)/P^\circ$ . If the Draize eye score ( $DES$ ) were related to a transport driven mechanism, the transfer process would be from the pure organic liquid into an initial biophase that will be the tear film and cell membranes on the surface of the eye. The more soluble the organic liquid in the initial phase, the larger the  $DES$  and hence greater irritation. Thus  $DES$  values would be proportional to  $1/g^\circ$ , the physicochemical solubility and hence  $\text{Log}(DES/P^\circ) = \text{Log } L$  where  $P^\circ$  is the saturated vapour pressure in ppm at 25°C. A general equation for the correlation and prediction of a series of  $\text{Log } L$  values for solutes into a given condensed phases had already been established.

$$\text{Log SP} = c + r R_2 + s\pi_2^H + a\alpha_2^H + b\beta_2^H + 1. \text{Log } L^{16} \quad (\text{Eq 6})$$

Application of Eq 6 to  $\text{Log}(DES)$  values yielded an extremely poor correlation but when  $\text{Log}(DES/P^\circ)$  was used as the dependent variable, a strong relationship (Eq 7) was found.

$$\text{Log}(DES/P^\circ) = -6.955 + 0.1046\pi_2^H + 4.437 \alpha_2^H + 1.350 \beta_2^H + 0.754 \text{Log } L^{16} \quad (\text{Eq 7})$$

$n = 37, r^2 = 0.951, SD = 0.32, F = 155.9$

On transforming the calculated  $\text{Log}(DES/P^\circ)$  values back to calculated  $DES$  values, there was good agreement with the original  $DES$  values (Eq 8).

$$\text{Log}(DES)_{\text{obs}} = 0.022 + 0.979 \text{Log}(DES)_{\text{calc}} \quad (\text{Eq 8})$$

$n = 37, r^2 = 0.771, SD = 0.3, F = 117.6$

It was suggested that the  $DES/P^\circ$  values referred to the transfer of the irritants from the vapour phase to the biophase and hence that a major factor in the Draize eye test was simply the transfer of the liquid (or the vapour) to the biological system.

25. Models for skin sensitisation have varied from those based on an *a priori* approach to those interpreted *a posteriori*. An example of both is described here. The first physicochemical mathematical model for skin sensitisation was the RAI (Relative Alkylation Index) model (124). This index quantifies the relative extent of sensitizer binding to the skin protein as a function of the dose given, the chemical reactivity (which could be expressed in the terms of the measured rate constants for reaction with a model nucleophile, in terms of Taft or Hammett substituent constants or in terms of computed molecular orbital indices) and hydrophobicity expressed as the octanol/water partition coefficient. The general form of the RAI expression is:

$$\text{RAI} = \text{Log } D + a \text{ Log } k + b \text{ Log } P \quad (\text{Eq 9})$$

where D is dose, k is the relative rate constant and P is the octanol/water partition coefficient. Log P here models both penetration and lipid/polar fluid partitioning.

26. Topological indices are often thought of as being difficult to interpret. In this example a model for skin sensitisation was developed relating the potency of a set of 93 diverse chemicals to a range of topological indices (125). The indices used in the final model accounted for hydrophobicity (H), polar surface area (PS), molar refractivity (MR), polarisability (PSR), charges (GM), van der Waals radii (VDW). Such parameters can be assigned as relevant in the context of skin sensitisation in that partition could be modelled by hydrophobicity, polar surface area, molar refractivity, van der Waals radii as bulk parameters and the reactivity accounted for by polarisability and charges. The Topological Sub-Structural Molecular Design (TOPS-MODE) approach used in this example is based on the method of moments (126, 127, 128). The approach consists of using the topological bond matrix (edge adjacency matrix) of the molecular graph. Bond weights in the main diagonal entries of the bond matrix are used to account for effects that could be involved in biological processes. An advantage with this approach is that a structural interpretation of TOPS-MODE results can be carried out by using the bond contributions to skin sensitization. These are calculated on the basis of the local moments which are defined as the diagonal entries of the different powers of the weighted bond matrix. This provides a mechanistic interpretation at a bond level and enables the generation of new hypotheses such as structural alerts.

### ***Environmental Endpoints***

27. The following (Q)SAR, taken from the European Technical Guidance Document for chemical risk assessment (129), predicts the acute toxicity of organic chemicals to the fathead minnow (*Pimephales promelas*). The equation developed was:

$$\text{Log (LC50)} = -0.846 \log K_{ow} - 1.39 \quad (\text{Eq 10})$$

where LC50 is the concentration (in moles per litre) causing 50% lethality in *Pimephales promelas*, after an exposure of 96 hours; and K<sub>ow</sub> is the octanol-water partition coefficient.

28. The (Q)SAR was developed for chemicals considered to act by a single mechanism of toxic action, non-polar narcosis, as defined by Verhaar *et al.* (130), and therefore has a clear mechanistic basis. In fact, non-polar narcosis is one of the most established mechanisms of toxic action. Non-polar narcosis has been established experimentally by using the Fish Acute Toxicity Syndrome methodology (131). The (Q)SAR is based on a descriptor for hydrophobicity (log K<sub>ow</sub>), which is relevant to the mechanism of action, i.e. toxicity results from the accumulation of molecules in biological membranes.

### ***Expert Systems***

29. An expert system for predicting toxicity is considered to be any formalised system not necessarily computer based, which enables a user to obtain rational predictions about the toxicity of chemicals. All expert systems for the prediction of chemical toxicity are built upon experimental data representing one or more manifestations of chemicals in biological systems

(the database) and/or rules derived from such data (the rulebase). Individual rules within the rulebase are generally of two main types. Some rules are based on mathematical induction whereas other rules are based on existing knowledge and expert judgement. Typically induced rules are QSARs whereas expert rules are often based on knowledge about reactive chemistry. Expert systems are sometimes characterized according to the nature of the rules in their rulebase. An expert system based primarily on statistically induced rules is sometimes called an “automated rule-induction system”, whereas a system based primarily on expert rules is referred to as a “knowledge based system” (132). The following two examples, referring to ECOSAR and DEREKfW, outline the mechanistic interpretation for these two types of expert system.

30. As part of the work by the OECD (Q)SAR Group, the ECOSAR tool was evaluated with respect to the OECD principles (5). ECOSAR (133, 134) predicts defined endpoints as required by the US EPA regulatory framework, such as acute L(E)C50 and long-term NOECs for fish, daphnids and algae. The (Q)SAR equations are based on linear regression analysis, using log Kow as the sole descriptor for predicting the L(E)C50 values (except for the class of surfactants). There is no explicit description of the chemical classes or the exclusion rules. The (Q)SAR for neutral organics is based on the assumption that all chemicals have a minimal toxicity based on the interference of the chemical with biological membranes, which can be modelled by the octanol-water partition coefficient (Kow). All other chemical classes show excess toxicity compared to the neutral organics.

31. DEREKfW is a knowledge-based expert system created with knowledge of structure-toxicity relationships and an emphasis on the need to understand mechanisms of action and metabolism. The DEREK knowledge base covers a broad range of toxicological endpoints, including mutagenicity, carcinogenicity and skin sensitisation.

32. The expert knowledge incorporated into the DEREKfW system originated from Sanderson & Earnshaw (135). These workers identified a series of ‘structural alerts’ associated with certain types of toxic activity. The DEREK knowledge base was written, developed and continues to be enhanced by LHASA (Logic and Heuristics Applied to Synthetic Analysis) Ltd and its members at the School of Chemistry, University of Leeds, UK. LHASA Ltd is a non-profit making collaboration consisting of the University of Leeds and various other educational and commercial institutions (including agrochemical, pharmaceutical and regulatory organizations) created to oversee the development of the DEREKfW system and the evolution of its toxicity knowledge base.

33. DEREKfW provides an explicit description of the substructure and substituents. When a query structure is processed, the alerts that match are displayed in a hierarchy called the prediction tree and are highlighted in bold in the query structure. The prediction tree includes the endpoint, the species and reasoning outcome, the number and name of the alert, and the example from the knowledge base if it exactly matches the query structure. The alert description provides a description depicting the structural requirement for the toxicophore detected and a reference to show the bibliographic references used. Some rules are extremely general with substructures only taking into account the immediate environment of a functional group. This means that remote fragments that may modulate a toxicity are not always taken into consideration. In other cases, the descriptions are much more specific.

34. All the rules in DEREK are based on either hypotheses relating to mechanisms of action of a chemical class or observed empirical relationships, the ideas for which come from

a variety of sources, including published data or suggestions from the DEREK collaborative group. This group consists of toxicologists who represent LHASA Ltd and customers who meet at regular intervals to give advice and guidance on the development of the databases and rulebases. The hypotheses underpinning each alert are documented in the alert descriptions as comments. These comments often include descriptions of features acting as electrophiles or nucleophiles. However, the detail depends on the specific alert. Some alerts contain no comments, apart from the modulating factors of skin penetration.

### ***Artificial Intelligence systems***

35. Many of the models so far discussed involve the use of transparent algorithms, typically regression equations where the mechanistic interpretation is achieved by interpreting the descriptors, the size of their coefficients, and perhaps the mathematical form of the equation. In contrast, AI-based models are sometimes considered to be non-transparent, since the algorithms are deeply embedded.

36. For example, Kohonen networks are specific types of networks that can provide mechanistic insights. Graphical representations of individual layers may indicate the roles of individual descriptors in the model. When a new compound is presented to the model it will be located on a defined position in the Kohonen network.. Its mechanism of activity may be deduced from the mechanisms of neighbouring compounds.

### **Concluding remarks**

37. There are many types of different types of modelling approaches. In this chapter, guidance is presented through the use of examples, to illustrate how to consider mechanism in the context of different types of model.

38. The mechanistic rationale of a (Q)SAR can be established *a priori*, in which case the descriptors are selected before modelling on the basis of their known or anticipated role in driving the response, or *a posteriori*, in which case the descriptors are selected on the basis of statistical fit alone, with their mechanistic rationale being rationalised after modelling. Models can also be developed by a combination of these two approaches.

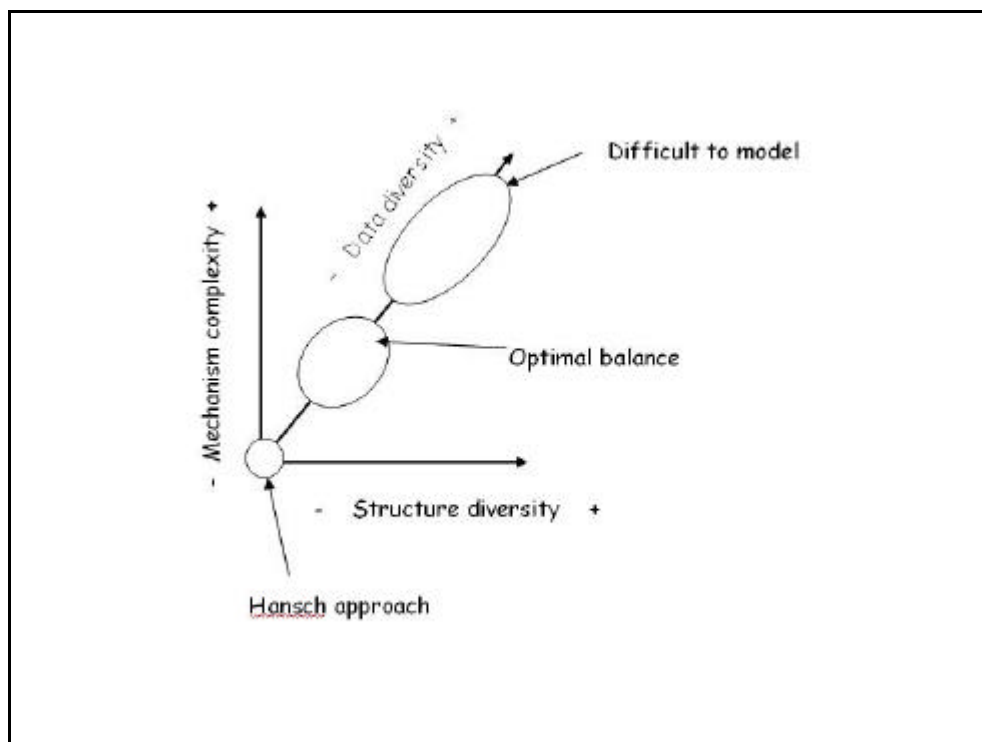
39. In the case of a QSAR with continuous descriptors, a mechanistic interpretation can be based on the physicochemical interpretation of each descriptor and its association with a mode or mechanism of action. The magnitudes of the model coefficients and model structure might also be taken into consideration.

40. In the case of a SAR, a mechanistic interpretation can be based on the chemical reactivity or molecular interaction of the substructure.

41. In the case of expert systems, it is not possible to generalise how a mechanistic interpretation could be assigned, due to the variety of such systems. Some systems are based primarily on expert knowledge, whereas others are based primarily on learned rules. For example, DEREK for Windows is based on the use of multiple structural alerts, each of which has its own scientific supporting evidence; whereas METEOR and CATABOL incorporate a significant amount of information on known metabolic pathways.

The architecture of neural network models does not generally correspond in any obvious way with underlying mechanisms of action.

**Figure 5.1. The Balance between Structural, Data and Mechanistic diversity**



**Table 5.1. Commonly used molecular descriptors in QSAR studies**

Molecular descriptor	Physicochemical interpretation	Examples of QSAR applications
<b>Logarithm of the Partition coefficient:</b> $\log P = \log (C_{\text{org}} / C_{\text{water}})$ $C_{\text{org}}$ = concentration of the non-ionised solute in the organic phase $C_{\text{water}}$ = concentration of the non-ionised solute in the water phase	Describes the distribution of a compound between organic (usually octanol) and water phase $\log P > 0$ – greater solubility in the organic phase; $\log P < 0$ – greater solubility in the aqueous phase. Measure of hydrophobicity / lipophilicity	Many applications in QSAR analysis of toxicological data sets (136)
<b>Hydrophobic substituent constant (p) :</b> $\pi_X = \log P_{R-X} - \log P_{R-H}$ $\log P_{R-H}$ = $\log P$ of the parent compound $\log P_{R-X}$ = $\log P$ of X substituted derivative	Describes the contribution of a substituent to the lipophilicity of a compound.	QSAR for mutagenicity of substituted N-nitroso-N-benzylmethyamines (137, 138)
<b>Hammett electronic substituent constant (s) :</b> $\log(K_{a_X}/K_{a_H}) = \rho \sigma$ $K_{a_H}$ = acid dissociation constant of benzoic acid $K_{a_X}$ = acid dissociation constant of X substituted derivative of benzoic acid $\rho$ = a series constant	Describes the electron-donating or -accepting properties of an aromatic substituent, in the <i>ortho</i> , <i>meta</i> and <i>para</i> positions.	QSARs of the relative toxicities of monoalkylated or monohalogenated benzyl alcohols (139)
<b>Taft steric parameter (<math>E_s</math>) :</b> $\log k = \log k_0 + r^* s^* + dE_s$ $s^*$ = polar substituent constant $r^*$ = constant	Steric substituent constant. Describes the intramolecular steric effects on the rate of a reaction.	Original reference of the formulation of Taft steric parameter (140)
<b>Aqueous solubility (<math>S_{aq}</math>) :</b> The maximum concentration of the compound that will dissolve in pure water at a certain temperature, at equilibrium	Measures the hydrophilicity of a compound	QSARs for fish bioconcentration factor (141)
<b>Molecular refractivity (MR):</b> $MR = [(n^2 - 1)/(n^2 + 2)] * M/\rho$ $n$ = refractive index $M$ = relative molecular mass $\rho$ = density	Describes the size and polarizability of a fragment or molecule. It could be considered as both an electronic and a steric parameter.	QSARs for binding of tetrahydroisoquinoline derivatives with estrogen receptors (142)
<b>Dissociation Constant (pKa)</b>	Describes extent of ionization of a compound. Reflects electron-directing effects of substituents.	QSARs for relative toxicity of monosubstituted phenols (143)



**Table 5.1. Commonly used molecular descriptors in QSAR studies**

<b>Dipole moment</b> Determined via experimental measurement of dielectric constant, refractive index and density, or calculated using molecular orbital theory	Describes separation of charge (polarity) in a molecule, and also considered as measure of hydrophilicity. Hypothesised to reflect the influence of electrostatic interactions with biological macromolecules (144)	QSARs for eye irritation of neutral organic chemicals (145)
<b>Atomic charge</b> Calculated by different molecular orbital methods	Descriptor that determines the electrostatic potential around a molecule, thus influencing intermolecular interactions with electrostatic nature.	QSARs for mutagenicity of quinolines (146)
<b>HOMO</b> (Highest Occupied Molecular Orbital) and <b>LUMO</b> (Lowest Unoccupied Molecular Orbital) reactivity indices. Calculated using molecular orbital theory.	Descriptors of molecular orbital energies. The HOMO energy describes the nucleophilicity of a molecule, whereas the LUMO energy describes electrophilicity.	Mutagenicity of aromatic and heteroaromatic amines (146, 147)
<b>Hydrogen bonding</b> Various measures have been proposed.	Descriptors of chemical reactivity (electrostatic interactions between molecules). Hydrogen-bond donors are proton donors (electronegative atoms or groups) and hydrogen-bond acceptors are groups with the capacity to donate a lone electron pair.	Modelling of aquatic toxicity of environmental pollutants (148)
<b>Molecular weight (MW) and Molecular volume (MV):</b> $MV = MW/\rho$ $\rho$ - density	Simple molecular size descriptors.	QSPR models for <i>in vivo</i> blood-brain partitioning of diverse organic compounds (149) QSARs of a series of xanthates as inhibitors and inactivators of cytochrome P450 2B1 (150)
<b>Molecular surface area (MSA)</b>	Size descriptor defined on the basis of the van der Waals surface of an energy minimised molecule by excluding gaps and crevices	Prediction of blood-brain partitioning for structurally diverse molecules (151)
<b>Topological Descriptors</b> Numerous types have been proposed, e.g. Wiener, Randic, Zagreb, Hosoya, Balaban, Kier and Hall molecular connectivity indices, kappa indices	Descriptors based on chemical graph theory, calculated from the connectivity tables of molecules.  Used to express different aspects of the shape and size of molecules, including degree of branching, and flexibility.	Modelling structural determinants of skin sensitisation (125, 152) QSAR of Phenol Toxicity (153)

**Table 5.1. Commonly used molecular descriptors in QSAR studies**

<b>Electrotopological descriptors</b>	Atom-based topological descriptors that encode information about the topological environment and electronic interactions of the atom.	QSAR Models for Antileukemic Potency of Carboquinones (154)
<b>Electronic Density Function (?)</b>	Descriptors of molecular similarity, based on electrostatic and steric interactions of the molecule	QSAR of antimycobacterial benzoxazines (155)
Obtained from Quantum Chemical Calculations.		

## REFERENCES

## REFERENCES

1. Jaworska, J.S., Comber, M., Auer, C. & van Leeuwen, C.J. (2003). Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. *Environmental Health Perspectives* **111**, 1358-1360.
2. Eriksson, L., Jaworska, J.S., Worth, A.P., Cronin, M.T.D., McDowell, R.M. & Gramatica, P. (2003). Methods for reliability, uncertainty assessment, and applicability evaluations of classification and regression based QSARs. *Environmental Health Perspectives* **111**, 1361-1375.
3. Cronin, M.T.D., Walker, J.D., Jaworska, J.S., Comber, M.H.I., Watts, C.D. & Worth, A.P. (2003). Use of quantitative structure-activity relationships in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. *Environmental Health Perspectives* **111**, 1376-1390.
4. Cronin, M.T.D., Jaworska, J.S., Walker, J.D., Comber, M.H.I, Watts, C.D. & Worth, A.P. (2003). Use of quantitative structure-activity relationships in international decision-making frameworks to predict health effects of chemical substances. *Environmental Health Perspectives* **111**, 1391-1401.
5. OECD. (2004). The Report from the Expert Group on (Quantitative) Structure-Activity Relationships [(Q)SARs] on the Principles for the Validation of (Q)SARs. ENV/JM/TG(2004)27/REV. Organisation for Economic Cooperation and Development, Paris, 17pp.
6. OECD. (2004). Annexes to the Report on the Principles for Establishing the Status of Development and Validation of (Quantitative) Structure-Activity Relationships [(Q)SARs]. ENV/JM/TG(2004)27/ANN. Organisation for Economic Cooperation and Development, Paris, 183pp.
7. OECD. (2005). OECD Guidance Document 34 on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment. ENV/JM/MONO(2005)14. Organisation for Economic Cooperation and Development, Paris, 96pp.
8. OECD. (2004). Regulatory Application of (Q)SARs: a U.S. EPA Case Study. ENV/JM/TG(2004)25/REV2. Organisation for Economic Cooperation and Development, Paris, 14pp.
9. OECD. (2005). Report on the Regulatory Uses and Applications in OECD Member Countries of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models in the Assessment of New and Existing Chemicals. Draft of September 2005, 56 pp.
10. Kier, L.B. & Hall, L.H. (1986). Molecular connectivity in structure-activity analysis. Res. Studio Press Ltd, Letchworth, UK.

11. Protic M. & Sabljic, A. (1989). Quantitative structure-activity relationships for acute toxicity of commercial chemicals on fathead minnows: Effect of molecular size. *Aquatic Toxicology* **14**, 47-64
12. Govers, H., Ruepert, C. & Aiking, H. (1984). Quantitative structure-activity relationships for polycyclic aromatic hydrocarbons: correlation between molecular connectivity, physico-chemical properties, bioconcentration and toxicity in *Daphnia pulex*. *Chemosphere* **13**, 227-236.
13. Sabljic, A. (1991). Chemical topology and ecotoxicology. *Science of the Total Environment* **109/110**, 197-220.
14. Verhaar, H.J.M., Rorije E., Borkent H., Seinen, W. & Hermens, J.L.M. (1996). Modelling the nucleophilic reactivity of organochlorine electrophiles: A mechanistically-based quantitative structure-activity relationship. *Environmental Toxicology and Chemistry* **16**, 1011-1018.
15. Purdy, R. (1991). The utility of computed superdelocalizability for predicting the LC50 values of epoxides to guppies. *Science of the Total Environment* **109/110**, 553-556.
16. Lewis, D.F.V. (1992). Computer-assisted methods in the evaluation of chemical toxicity. In: Reviews in Computational Chemistry, Vol. III (Lipkowitz, K.B. and Boyd, D.B., Eds), pp. 173-222. VHC Publishers, New York.
17. Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T.D., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D.W., Schultz, T.W., Stanton, D.T., van de Sandt, J.J.M, Tong, W., Veith, G. & Yang, C. (2005). Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. *Alternatives to Laboratory Animals* **33**, 155- 173.
18. Deneer, J.W., Seinen, W. & Hermens, J.L.M. (1988). The acute toxicity of aldehydes to guppy. *Aquatic Toxicology* **12**, 185- 192.
19. Schultz, T.W. & Cronin, M.T.D. (1999). Response-surface analysis for toxicity to *Tetrahymena pyriformis*: reactive carbonyl-containing chemicals. *Journal of Chemical Information and Computer Sciences* **39**, 304- 309.
20. Verhaar, H.J.M., Mulder, W. & Hermens, J.L.M. (1995). QSARs for ecotoxicity. In: Overview of structure-activity relationships for environmental endpoints. Part 1: General outline and procedure. Hermens, J.L.M. (Ed.). Report prepared within the framework of the project "QSAR for Prediction of Fate and Effects of Chemicals in the Environment". Contract with the European Commission EV5V-CT92-0211.
21. Schultz, T.W., Cronin, M.T.D., Netzeva, T.I. & Aptula, A.O. (2002). Structure-toxicity relationships for aliphatic chemicals evaluated with *Tetrahymena pyriformis*. *Chemical Research in Toxicology* **15**, 1602–1609.
22. Russom, C.L., Bradbury, S.P., Broderius, S.J., Hammermeister, D.E. & Drummond, R.A. (1997). Predicting modes of toxic action from chemical structure: acute toxicity in

- the fathead minnow (*Pimephales promelas*). *Environmental Toxicology and Chemistry* **16**, 948- 967.
23. Schultz, T.W., Netzeva, T.I., Roberts, D.W. & Cronin, M. T. D. (2005). Structure-toxicity relationships for the effects to *Tetrahymena pyriformis* of aliphatic, carbonyl-containing  $\alpha,\beta$ -unsaturated chemicals. *Chemical Research in Toxicology* **18**, 330- 341.
  24. O'Brien, P.J. (1991). Molecular mechanisms of quinone toxicity, *Chemico-Biological Interactions* **80**, 1- 41.
  25. Jaworska, J., Aldenberg, T. & Nikolova, N. (2005). Review of methods for assessing the applicability domain of SARs and QSARs. *Alternatives to Laboratory Animals*, in press.
  26. Gramatica, P., Pilutti, P. & Papa, E. (2004). Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling. *Journal of Chemical Information and Computer Sciences* **44**, 1794- 1802.
  27. Pavan, M., Worth, A. & Netzeva, T. (2005). Preliminary analysis of an aquatic toxicity dataset and assessment of QSAR models for narcosis. JRC report EUR 21479 EN. European Commission, Joint Research Centre, Ispra, Italy.
  28. Jaworska, J.S., Aldenberg, T. & Nikolova, N. (2005). Review of methods for assessing the applicability domains of SARs and QSARs. *Alternatives to Laboratory Animals*, in press.
  29. Ambit Disclosure software developed by Jaworksa, J.S. and Nikolova, N. Accessible: <http://ambit.acad.bg/>
  30. Tong, W., Hong, H., Fang, H., Xie, Q. & Perkins, R. (2003). Decision forest: combining the predictions of multiple independent decision tree models. *Journal of Chemical Information and Computer Sciences* **43**, 525- 531.
  31. Dimitrov, S., Dimitrova, G., Pavlov, T., Dimitrova, N., Patlewicz, G., Niemela, J. & Mekenyan, O. (2005) A stepwise approach for defining the applicability domain of SAR and QSAR models. *Journal of Chemical Information and Modeling* **45**, 839 -849.
  32. Cunningham, A.R. & Rosenkranz, H.S. (2001). Estimating the extent of the health hazard posed by high-production volume chemicals. *Environmental Health Perspectives* **110**, 953- 956.
  33. Klopman, G., Chakravarti, S.K., Harris, N., Ivanov, J. & Saiakov, R.D. (2003). In-silico screening of high production volume chemicals for mutagenicity using the MCASE QSAR expert system. *SAR and QSAR in Environmental Research* **14**, 165- 180.
  34. Hong, H., Tong, W., Fang, H., Shi, L., Xie, Q., Wu, J., Perkins, R., Walker, J.D., Branham, W. & Sheehan, D. (2002). Prediction of estrogen receptor binding for 58,000 chemicals using an integrated system of a tree-based model with structural alerts. *Environmental Health Perspectives* **110**, 29- 36.

35. Schmieder, P., Mekenyan, O., Bradbury, S. & Veith, G. (2003). QSAR Prioritisation of chemical inventories for endocrine disruptor testing. *Pure and Applied Chemistry* **75**, 2389- 2396.
36. Tunkel, J., Mayo, K., Austin, C., Hickerson, A. & Howard, P. (2005). Practical Considerations on the use of predictive models for regulatory purposes. *Environmental Science and Technology* **39**, 2188- 2199.
37. Netzeva, T.I., Gallegos Saliner, A. & Worth, A.P. (2005). Comparison of the applicability domain of a QSAR for estrogenicity with a large chemical inventory. *Environmental Toxicology and Chemistry*, in press.
38. Jouan-Rimbaud, D., Massart, D.L. & de Noord, O.E. (1996). Random Correlation in Variable Selection for Multivariate Calibration with a Genetic Algorithm. *Chemometrics and Intelligent Laboratory Systems* **35**, 213-220.
39. Hawkins, D.M. (2004). The Problem of Overfitting. *Journal of Chemical Information & Computer Sciences* **44**, 1-12.
40. Topliss, J.G. & Edwards, R.P. (1979). Chance Factors in Studies of Quantitative Structure-Activity Relationships. *Journal of Medicinal Chemistry* **22**, 1238-1244.
41. Wold, S. & Dunn III, W.J. (1983). Multivariate Quantitative Structure-Activity Relationships (QSAR): Conditions for their Applicability. *Journal of Chemical Information & Computer Sciences* **23**, 6-13.
42. Clark, M. & Cramer III, R.D. (1993). The Probability of Chance Correlation Using Partial Least Squares (PLS). *Quantitative Structure-Activity Relationships* **12**, 137-145.
43. Massart, D.L., Vandeginste, B.G., Buydens, L.M., Lewi, P.J. & Smeyers-Verbeke, J. (1997). Handbook of Chemometrics and Qualimetrics: Part A, Elsevier Science.
44. Kubinyi, H. (1993). QSAR: Hansch Analysis and Related Approaches. Methods and Principles in Medicinal Chemistry, Vol. 1 (Mannhold, R., Kroogsgard-Larsen, P. & Timmerman, H., Eds.). VCH, Weinheim.
45. Wold, S., Ruhe, A., Wold, H. & Dunn III, W. J. (1984). The collinearity problem in linear regression, the partial least squares (PLS) approach to generalized inverses. *SIAM Journal of Science Statistics and Computer* **5**, 735-743.
46. Wold, S., Johansson, E. & Cocchi, M. (1993). PLS: Partial least squares projections to latent structures. In: 3D-QSAR in Drug Design: Theory, Methods and Applications (Kubinyi, H., Ed.), pp523-550. ESCOM Science, Leiden, The Netherlands.
47. Wold, S. (1995). PLS for Multivariate Linear Modeling, in Chemometric Methods in Molecular Design (van de Waterbeemd, H., Ed.). VCH, Weinheim, Germany.
48. Massart, D.L., Vandeginste, B.G., Buydens, L.M., Lewi, P.J. & Smeyers-Verbeke, J. (1997). Handbook of Chemometrics and Qualimetrics: Part B. Elsevier Science, Amsterdam, The Netherlands.

49. Eriksson, L., Johansson, E., Kettaneh-Wold, N. & Wold, S. (2001) Multi- and Megavariate Data Analysis. Principles and Applications. Umetrics AB, Umeå, Sweden.
50. Netzeva, T.I., Schultz, T.W., Aptula, A.O. & Cronin, M.T.D. (2003). Partial least squares modelling of the acute toxicity of aliphatic compounds to *Tetrahymena pyriformis*. *SAR and QSAR in Environmental Research* **14**, 265-283.
51. Eriksson, L., Jaworska, J., Worth, A., Cronin, M.T.D., McDowell, R.M. & Gramatica, P. (2003). Methods for reliability, uncertainty assessment, and applicability evaluations of regression based and classification QSARs. *Environmental Health Perspectives* **111**, 1361-1375.
52. Worth, A.P & Cronin, M.T.D. (2003). The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. *Journal of Molecular Structure (Theochem)* **622**, 97-111.
53. Worth, A.P. & Cronin, M.T.D. (2000). Embedded Cluster Modelling: A Novel Quantitative Structure-Activity Relationship for Generating Elliptic Models of Biological Activity. In: Progress in the Reduction, Refinement and Replacement of Animal Experimentation (Balls, M., van Zeller, A.M. & Halder, M.E., Eds.), pp. 479 – 491. Elsevier Science, Amsterdam, The Netherlands.
54. Cooper, J.A., Saracci, R., & Cole, P. (1979). Describing the validity of carcinogen screening tests. *British Journal of Cancer* **39**, 87–89.
55. Feinstein, A.R. (1975). Clinical biostatistics XXXI. On the sensitivity, specificity, and discrimination of diagnostic tests. *Clinical Pharmacology & Therapeutics* **17**, 104–116.
56. Sullivan Pepe, M. (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford Statistical Science Series **28**. Oxford University Press.
57. McDowell, R.M. & Jaworska, J. (2002). Bayesian analysis and inference of QSAR predictive model results. *SAR and QSAR in Environmental Research* **13**, 111–125.
58. Frank, I.E. & Friedman, J.H. (1989). Classification: oldtimers and newcomers. *Journal of Chemometrics* **3**, 463-475.
59. Kraemer, H.C. (1982). Kappa coefficient. In: Encyclopedia of Statistical Sciences (Kotz, S. & Johnson, N. L. ,Eds.). John Wiley & Sons, New York.
60. Wehrens, R., Putter, H. & Buydens, L.M.C. (2000). Bootstrap: a tutorial. *Chemometrics and Intelligent Laboratory Systems* **54**, 35-52.
61. Worth, A.P. & Cronin, M.T.D. (2001). The use of bootstrap resampling to assess the uncertainty of Cooper statistics. *Alternatives to Laboratory Animals* **29**, 447-459.
62. Lusted, L.B. (1971). Signal detectability and medical decision-making. *Science* **171**, 1217-1219.



63. Hanley, J.A. (1989). Receiver operating characteristic (ROC) methodology: The state of the art. *Critical Reviews in Diagnostic Imaging* **29**, 307-335.
64. Provost, F. & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning Journal* **42**, 203 – 231.
65. Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). Classification and Regression Trees. Wadsworth International Group, Belmont, CA, USA.
66. Hand, D. (1981). Discrimination and Classification. Wiley & Sons, New York.
67. Lek, S. & Guegan, J.F. (1999). Artificial neural networks as a tool in ecological modeling, an introduction. *Ecological Modelling* **120**, 65-73.
68. Zupan, J. & Gasteiger, J. (1999). Neural Networks for Chemistry and Drug Design, Wiley-VCH, Weinheim (Germany).
69. Anzali, S., Gasteiger, J., Holzgrabe, U., Polanski, J., Sadowski, J., Teckentrup, A., & Wagener, M. (1998). The Use of Self-Organizing Neural Networks in Drug Design.
70. Kubinyi, H., Folkers, G., Martin, & Y.C. (1998). 3D QSAR in Drug Design - Vol. 2. Kluwer Academic, New York.
71. Spycher, S., Pellegrini, E. & Gasteiger, J. (2005). Use of structure descriptors to discriminate between modes of toxic action of phenols. *Journal of Chemical Information and Modeling* **45**, 200-208.
72. Vracko, M. (2005). Kohonen artificial neural network and counter propagation neural network in molecular structure-toxicity studies. *Current Computer-Aided Drug Design* **1**, 73-78.
73. Guha, R. & Jurs, P.C. (2005). Determining the validity of a QSAR model – a classification approach. *Journal of Chemical Information and Modeling* **43**, 65-73.
74. Mazzatorta, P., Vracko, M., Jezierska, A. & Benfenati, E. (2003). Modeling toxicity by using supervised Kohonen neural networks. *Journal of Chemical Information and Modeling* **43**, 485-492.
75. Vracko, M. & Gasteiger, J. (2002). A QSAR study on a set of 105 flavonoid derivatives using descriptors derived from 3D structures. *Internet Electronic Journal of Molecular Design* **1**, 527-544.
76. Devillers, J. & Domine, D. (1999). A noncongeneric model for predicting toxicity of organic molecules to vibrio fischeri. *SAR & QSAR in Environmental Research* **10**, 61-70.
77. Diaconis, P. & Efron, B. (1983). Computer Intensive Methods in Statistics. *Scientific American* **248**, 96-108.

78. Cramer III, R.D., Bunce, J.D., Patterson, D.E. & Frank, I.E. (1988). Cross validation, bootstrapping and Partial Least Squares compared with multiple regression in conventional QSARsStudies. *Quantitative Structure-Activity Relationships* **7**, 18-25.
79. Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of American Statistical Association* **78**, 316-331.
80. Osten, D.W. (1988). Selection of optimal regression models via cross-validation. *Journal of Chemometrics* **2**, 39-48.
81. Cruciani, G., Baroni, M., Clementi, S., Costantino, G., Riganelli, D. & Skagerberg, B. (1992). Predictive ability of regression models. Part I: standard deviation of prediction errors (SDEP). *Journal of Chemometrics* **6**, 335-346.
82. Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, London.
83. Lindgren, F., Hansen, B., Karcher, W., Sjöström, M. & Eriksson, L. (1996). Model validation by permutation tests: applications to variable selection. *Journal of Chemometrics* **10**, 421-532.
84. Todeschini, R., Consonni, V. & Maiocchi, A. (1999). The K correlation index: theory development and its applications in chemometrics. *Chemometrics and Intelligent Laboratory Systems* **46**, 13-29.
85. Todeschini, R., Consonni, V., Mauri, A. & Pavan, M. (2004). Detecting “bad” regression models: multicriteria fitness functions in regression analysis. *Analytica Chimica Acta* **515**, 199-208.
86. Stone, M. & Jonathan, P. (1993). Statistical thinking and technique for QSAR and related studies. Part I: General theory. *Journal of Chemometrics* **7**, 455-475.
87. Mager, P.P. (1995). Diagnostics statistics in QSAR. *Journal of Chemometrics* **9** 211-221.
88. Boggia, R., Forina, M., Fossa, P. & Mosti, L. (1997). Chemometric study and validation strategies in the structureaActivity relationship of new cardiotonic agents. *Quantitative Structure-Activity Relationships* **16**, 201-213.
89. Golbraikh, A. & Tropsha, A. (2002). Beware of  $q^2$ ! *Journal of Molecular Graphics and Modelling* **20**, 269-276.
90. Golbraikh, A. & Tropsha, A. (2002). Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Journal of Computer Aided Molecular Design* **16**, 357-369.
91. Box, G.E.P., Hunter, W.G. & Hunter, J.S. (1978). *Statistics for Experimenters. An Introduction to Design, Data Analysis and Model Building*. John Wiley & Sons, USA

92. Tropsha, A., Gramatica, P. & Gombar, V.K. (2003). The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR & Combinatorial Science* **22**, 69-77.
93. Yasri, A. & Hartsough, D. (2001). Toward an optimal procedure for variable Selection and QSAR model building. *Journal of Chemical Information & Computer Sciences* **41**, 1218-1227.
94. Potter, T. & Matter, H. (1998). Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *Medicinal Chemistry* **41**, 478-488.
95. Taylor, R. (1995). Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *Journal of Chemical Information & Computer Sciences* **35**, 59-67.
96. Gastaiger, J. & Zupan, J. (1993). Neural networks in chemistry. *Angewandte Chemie International Edition* **32**, 503-527.
97. Kennard, R.W. & Stone, L.A. (1969). Computer aided design of experiments. *Technometrics* **11**, 137-148.
98. Eriksson, L. & Johansson, E. (1996). Multivariate design and modeling in QSAR. Tutorial. *Chemometrics and Intelligent Laboratory Systems* **34**, 1-19.
99. Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y., Lee, K. & Tropsha, A. (2003). Rational selection of training and test sets for the development of validated QSAR models. *Journal of Computer Aided Molecular Design* **17**, 241-253.
100. Kauffman, G.W. & Jurs, P.C. (2001). QSAR and *k*-Nearest Neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. *Journal of Chemical Information & Computer Sciences* **41**, 1553-1560.
101. Burden, F.R. (1999). Robust QSAR models using Bayesian regularized neural networks. *Journal of Medicinal Chemistry* **42**, 3183-3187.
102. Loukas, Y.L. (2001). Adaptive neuro-fuzzy inference system: an instant and architecture-free predictor for improved QSAR studies. *Journal of Medicinal Chemistry* **44**, 2772-2783.
103. Bourguignon, B., de Agular, P.F., Khots, M.S. & Massart, D.L. (1994). Optimization in irregularly shaped regions: pH and solvent strength in reversed phase High-Performance Liquid Chromatography separations. *Analytical Chemistry* **66**, 893-904.
104. Hudson, B.D., Hyde, R.M., Rahr, E. & Wood, J. (1996). Parameter based methods for compounds selection from chemical databases. *Quantitative Structure-Activity Relationships* **15**, 285-289.
105. Snarey, M., Terret, N.K., Willet, P. & Wilton, D.J. (1997). Comparison of algorithms for dissimilarity-based compound selection. *Journal of Molecular Graphics and Modeling* **15**, 373-385.

106. Nilakatan, R., Bauman, N. & Haraki, K.S. (1997). Database diversity assessment: New ideas, concepts and tools. *Journal of Computer Aided Molecular Design* **11**, 447-452.
107. Gobbi, A., & Lee, M-L. (2003). Database DISE: Directed Sphere Exclusion. *Journal of Chemical Information & Computer Sciences* **43**, 317-323.
108. Hammett, L.P. (1937). The effect of structure on the reaction of organic compounds. Benzene derivatives. *Journal of the American Chemical Society* **59**, 96-103.
109. Hansch, C, Maloney, P.P., Fujitya, T. & Muir, R.M. (1962). Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **194**, 178-180.
110. Hansch, C. & Fukita, T. (1964) Rho-sigma-pi analysis. A method for the correlation of biological activity with chemical structure. *Journal of the American Chemical Society* **86**, 1616-1626.
111. Hammett, L.P. (1970). Physical Organic Chemistry. Mc Graw Hill, New York.
112. Taft, R.W. (1956). Steric effects in Organic Chemistry. Wiley, New York.
113. Hansch, C., Smith, R.N., Rockoff, A., Calef, D.F., Jow, P.Y.C. & Fukunaga, J.Y. (1977). Structure-activity relationships in papain and bromelain ligand interactions *Archives of Biochemistry and Biophysics* **183**, 383-392.
114. Livingstone, D.L. (1995). Data Analysis for Chemists. Oxford Science Publications.
115. Benigni, R., Andreoli, C. & Giuliani, A. (1994). (Q)SAR models for both mutagenic potency and activity: application to nitroarenes and aromatic amines. *Environmental and Molecular Mutagenesis* **24**, 208-219.
116. Debnath, A. K., Debnath, G., Shusterman, A.J. & Hansch, C. (1992). A QSAR Investigation of the Role of Hydrophobicity in Regulating Mutagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in Salmonella Typhimurium TA98 and TA100. *Environmental and Molecular Mutagenesis* **19**, 37-52.
117. Barratt, M.D. (1995). A quantitative structure-activity relationship for the eye irritation potential neutral organic chemicals. *Toxicology Letters* **80**, 69-74.
118. EEC (1984). 84/449/EEC. Commission Directive of 25 April 1984 adapting to technical progress for the sixth time Council Directive 67/548/EEC on the approximation of laws, regulations and administrative procedures relating to the classification, packaging and labelling of dangerous substances. *Official Journal of European Communities* L25, 106-108.
119. OECD (1987). OECD Guidelines for Testing of Chemicals. Test Guideline 405, Acute Eye Irritation/Corrosion.

120. Bagley, D.M., Botham, P.A., Gardner, J.R., Holland, G., Kreiling, R., Lewis, R.W., Stringer, D.A. & Walker, A.P. (1992). Eye irritation: reference chemicals databank. *Toxicology in Vitro* **6**, 487-491.
121. Jacobs, G.A. & Martens, M.A. (1989). An objective method for the evaluation of eye irritation in vivo. *Food and Chemical Toxicology* **27**, 255-258.
122. Abraham, M.H., Kumarsingh, R., Cometto-Muniz, J.E. & Cain, W.S. (1998). A (Q)SAR for a Draize eye irritation database. *Toxicology in Vitro* **12**, 201-207.
123. Abraham M.H. (1994). Scales of solute hydrogen-bonding: their construction and application to physicochemical and biochemical processes. *Chemical Society Reviews* **22**, 73-83.
124. Roberts, D.W. & Williams, D.L. (1982). The derivation of quantitative correlations between skin sensitisation and physico-chemical parameters for alkylating agents and their application to experimental data for sulfones. *Journal of Theoretical Biology* **99**, 807-825.
125. Estrada, E., Patlewicz, G., Chamberlain, M., Basketter, D. & Larbey, S. (2003). Computer-aided knowledge generation for understanding skin sensitization mechanisms: The TOPS-MODE approach. *Chemical Research in Toxicology* **16**, 1226-1235.
126. Estrada, E. (1996). Spectral moments of the edge adjacency matrix in molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes. *Journal of Chemical Information and Computer Science* **36**, 844-849.
127. Estrada, E. (1997). Spectral moments of the edge-adjacency matrix of molecular graphs. 2. Molecules containing heteroatoms and (Q)SAR applications. *Journal of Chemical Information and Computer Science* **37**, 320-328.
128. Estrada, E. (1998). Spectral moments of the edge-adjacency matrix of molecular graphs. 3. Molecules containing cycles. *Journal of Chemical Information and Computer Science* **38**, 23-27.
129. EC (1996) Technical Guidance Document in Support of Commission Directive 93/67/EEC on Risk Assessment for New Notified Substances and Commission Regulation (EC) No 1488/94 on Risk Assessment for Existing Substances, Luxembourg: European Commission, Office for Official Publications of the European Communities.
130. Verhaar, H.J.M., van Leeuwen, C.J. & Hermens, J.L.M. (1992). Classifying environmental pollutants. 1. Structure-activity relationships for prediction of aquatic toxicity. *Chemosphere* **25**, 471-491.
131. McKim, J.M., Schmieder, P.K., Carlson, R.W., Hunt, E.P. & Niemi, G.I. (1987). Use of respiratorycardiovascular responses of rainbow-trout (*Salmo gairdneri*) in identifying acute toxicity syndromes in fish. 1. Pentachlorophenol, 2,4-dinitrophenol, tricaine methanesulfonate and 1-octanol, *Environmental Toxicology and Chemistry* **6**, 295-312.

132. Dearden, J.C., Barratt, M.D., Benigni, R. Bristol, D.W., Combes, R.D., Cronin, M.T.D. Judson, P.N., Payne, M.P., Richard, A.M., Tichy, M., Worth, A.P & Yourick, J.J. (1997). The development and validation of expert systems for predicting toxicity. *Alternatives to Laboratory Animals* **25**, 223-252.
133. ECOSAR (1996). Technical reference manual. See website: <http://www.epa.gov/oppt/newchems/sarman.pdf>
134. ECOSAR User manual (1998). See website: <http://www.epa.gov/oppt/newchems/manual.pdf>.
135. Sanderson, D.M. & Earnshaw, C.G. (1991). Computer prediction of possible toxic action from chemical structure. The DEREK system. *Human & Experimental Toxicology* **10**, 261-273.
136. Cronin, M.T.D, Dearden, J.C., Duffy, J.C., Edwards, R., Manga, N., Worth, A.P. & Worgan, A.D. (2002). The importance of hydrophobicity and electrophilicity descriptors in mechanistically-based QSARs for toxicological endpoints. *SAR & QSAR in Environmental Research* **13**, 167-176.
137. Singer, G.M., Andrews, A.W. & Guo, S.M. (1986). Quantitative structure-activity relationship of the mutagenicity of substituted N-nitroso-N-benzylmethyamines: possible implications for carcinogenicity. *Journal of Medicinal Chemistry* **29**, 40-44.
138. Benigni, R. (2005). Structure-activity relationship studies of chemical mutagens and carcinogens: mechanistic investigations and prediction approaches. *Chemical Reviews* **105**, 1767-1800.
139. Schultz, T.W., Jain, R., Cajina-Quezada, M. & Lin, D.T. (1988). Structure-toxicity relationships for selected benzyl alcohols and the polar narcosis mechanism of toxicity. *Ecotoxicology and Environmental Safety* **16**, 57-64.
140. Taft, R.W. (1956). Separation of Polar, Steric and Resonance Effects in reactivity. In Steric effects in Organic Chemistry. Newman, M.S. (Ed.). pp. 556-675. Wiley, New York,
141. Dearden, J.C. & Shinnawei, N.M. (2004). Improved prediction of fish bioconcentration factor of hydrophobic chemicals. *SAR & QSAR in Environmental Research* **15**, 449-455.
142. Hansch, C., Bonavida, B., Jazirehi, A.R., Cohen, J.J., Milliron, C. & Kurup, A. (2003). Quantitative structure-activity relationships of phenolic compounds causing apoptosis. *Bioorganic & Medicinal Chemistry* **11**, 617-620.
143. Schultz, T.W., Lin, D.T., & Wesley, S.K. (1992). QSARs for monosubstituted phenols and the polar narcosis mechanism of toxicity. *Quality Assurance* **1**, 132-143.
144. Dearden, J.C. (1990). Physico-chemical descriptors, in Practical applications of quantitative structure-activity relationships (QSAR) in environmental chemistry and toxicology (Karcher, W. & Devillers Eds), pp 25-61. Kluwer Academic Publishers.

145. Barratt, M.D. (1995). A quantitative structure-activity relationship for the eye irritation potential of neutral organic chemicals. *Toxicology Letters* **80**, 69-74.
146. Debnath, A.K., de Compadre, R.L. & Hansch, C. (1992). Mutagenicity of quinolines in *Salmonella typhimurium* TA100. A QSAR study based on hydrophobicity and molecular orbital determinants. *Mutation Research* **280**, 55-65.
147. Debnath, A. K., Debnath, G., Shusterman, A.J. & Hansch, C. (1992). A QSAR Investigation of the Role of Hydrophobicity in Regulating Mutagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in *Salmonella Typhimurium* TA98 and TA100. *Environmental and Molecular Mutagenesis* **19**, 37-52.
148. Raevsky, O.A., Dearden, J.C. (2004). Creation of predictive models of aquatic toxicity of environmental pollutants with different mechanisms of action on the basis of molecular similarity and HYBOT descriptors. *SAR & QSAR in Environmental Research* **15**, 433-448.
149. Hou, T.J. & Xu, X.J. (2003). ADME evaluation in drug discovery. 3. Modeling blood-brain barrier partitioning using simple molecular descriptors. *Journal of Chemical Information and Computer Sciences* **43**, 2137-2152.
150. Lesigiarska, I., Pajeva, I. & Yanev, S. (2002). Quantitative structure-activity relationship (QSAR) and three-dimensional QSAR analysis of a series of xanthates as inhibitors and inactivators of cytochrome P450 2B1. *Xenobiotica* **32**, 1063-1077.
151. Kaznessis, Y.N., Snow, M.E. & Blankley, C.J. (2001). Prediction of blood-brain partitioning using Monte Carlo simulations of molecules in water. *Journal of Computer-Aided Molecular Design* **15**, 697-708.
152. Estrada, E., Patlewicz, G., Chamberlain, M., Basketter, D. & Larbey, S. (2003). Computer-aided knowledge generation for understanding skin sensitization mechanisms: the TOPS-MODE approach. *Chemical Research in Toxicology* **16**, 1226-1235.
153. Hall, L. H. & Vaughn, T.A. (1997). QSAR of phenol toxicity using electrotopological state and kappa shape Indices. *Medicinal Chemistry Research* **7**, 407-416.
154. Gough, J. & Hall, L.H. (1999). QSAR Models of the antileukemic potency of carboquinones: electrotopological state and Chi Indices. *Journal of Chemical Information and Computer Science* **39**, 356-361.
155. Gallegos, A., Carbó-Dorca, R., Ponec, R. & Waisser, K. (2004) Similarity approach to QSAR. Application to antimycobacterial benzoxazines. *International Journal of Pharmaceutics* **269**, 51-60.





**APPENDIX 1**

**CHECKLIST FOR THE INTERPRETATION OF  
THE OECD (Q)SAR VALIDATION PRINCIPLES**

## **Introduction**

According to the OECD principles for (Q)SAR validation, a (Q)SAR should be associated with the following information:

- 1) a defined endpoint
- 2) an unambiguous algorithm
- 3) a defined domain applicability
- 4) appropriate measures of goodness-of-fit, robustness and predictivity
- 5) a mechanistic interpretation, if possible

This Appendix provides a series of questions associated with each principle, intended to provide an overview of the main considerations associated with the application of each principle. The questions are neither intended to be definitive, nor equally relevant for a given type of model.

## CHECK LIST FOR PROVIDING GUIDANCE ON THE INTERPRETATION OF THE OECD PRINCIPLES FOR (Q)SAR VALIDATION

PRINCIPLE	CONSIDERATIONS
	Is the following information available for the model ? <span style="float: right;">Yes/No/NA</span>
<b>1) Defined endpoint</b> 1.1 A clear definition of the scientific purpose of the model (i.e. does it make predictions of a clearly defined physicochemical, biological or environmental endpoint) ? 1.2 The potential of the model to address (or partially address) a clearly defined regulatory need (i.e. does it make predictions of a specific endpoint associated with a specific test method or test guideline) ? 1.3 Important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period, protocol) ? 1.4 The units of measurement of the endpoint ?	
<b>2) Defined algorithm</b> 2.1 In the case of a SAR, an explicit description of the substructure, including an explicit identification of its substituents ? 2.2 In the case of a QSAR, an explicit definition of the equation, including definitions of all descriptors ?	
<b>3) Defined domain of applicability</b> 3.1 In the case of a SAR, a description of any limits on its applicability (e.g. inclusion and/or exclusion rules regarding the chemical classes to which the substructure is applicable) ? 3.2 In the case of a SAR, rules describing the modulatory effects of the substructure's molecular environment ? 3.3 In the case of a QSAR, inclusion and/or exclusion rules that define the following variable ranges for which the QSAR is applicable (i.e. makes reliable estimates): a) descriptor variables ? b) response variables ? 3.4 A (graphical) expression of how the descriptor values of the chemicals in the training set are distributed in relation to the endpoint values predicted by the model ?	

#### **4A) Internal performance**

- 4.1 Full details of the training set given, including details of:
- a) number of training structures
  - b) chemical names
  - c) structural formulae
  - d) CAS numbers
  - e) data for all descriptor variables
  - f) data for all response variables
  - g) an indication of the quality of the training data ?
- 4.2
- a) An indication whether the data used to develop the model were based upon the processing of raw data (e.g. the averaging of replicate values)
  - b) If yes to a), are the raw data provided ?
  - c) If yes to a), is the data processing method described ?
- 4.3 An explanation of the approach used to select the descriptors, including:
- a) the approach used to select the initial set of descriptors
  - b) the initial number of descriptors considered
  - c) the approach used to select a smaller, final set of descriptors from a larger, initial set
  - d) the final number of descriptors included in the model ?
- 4.4
- a) A specification of the statistical method(s) used to develop the model (including details of any software packages used)
  - b) If yes to a), an indication whether the model has been independently confirmed (i.e. that the independent application of the described statistical method to the training set results in the same model) ?
- 4.5 Basic statistics for the goodness-of-fit of the model to its training set (e.g.  $r^2$  values and standard error of the estimate in the case of regression models) ?
- 4.6
- a) An indication whether cross-validation or resampling was performed
  - b) If yes to a), are cross-validated statistics provided, and by which method ?
  - c) If yes to a), is the resampling method described ?
- 4.7 An assessment of the internal performance of the model in relation to the quality of the training set, and/or the known variability in the response ?

#### **4B) Predictivity**

- 4.8 An indication whether the model has been validated by using a test set that is independent of the training set ?
- 4.9 If an external validation has been performed (yes to 4.8), full details of the test set, including details of:
- a) number of test structures
  - b) chemical names
  - c) structural formulae
  - d) CAS numbers
  - e) data for all descriptor variables
  - f) data for all response variables
  - g) an indication of the quality of the test data ?

- 4.10 If an external validation has been performed (yes to 4.8):
- a) an explanation of the approach used to select the test structures, including a specification of how the applicability domain of the model is represented by the test set ?
  - b) a specification of the statistical method(s) used to assess the predictive performance of the model (including details of any software packages used)
  - c) a statistical analysis of the predictive performance of the model (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models)
  - d) an evaluation of the predictive performance of the model that takes into account the quality of the training and test sets, and/or the known variability in the response
  - e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria ?

#### **5) Mechanistic interpretation**

- 5.1 In the case of a SAR, a description of the molecular events that underlie the properties of molecules containing the substructure (e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region) ?
- 5.2 In the case of a QSAR, a physicochemical interpretation of the descriptors that is consistent with a known mechanism of (biological) action ?
- 5.3 Literature references that support the (purported) mechanistic basis ?
- 5.4 An indication whether the mechanistic basis of the model was determined *a priori* (i.e. before modelling, by ensuring that the initial set of training structures and/or descriptors were selected to fit a pre-defined mechanism of action) or *a posteriori* (i.e. after the modelling, by interpretation of the final set of training structures and/or descriptors) ?